A Human-Computer Collaborative Video Creation Paradigm Integrating Creative Intent Memory, Preference Feedback, and Editable Generation Control

Xiaohan Wang[1], Xukun Wu[1*], Ying Liu[1], Muya Huang[1]
[1]School of Art and Design, Wanjiang University of Technology, Ma'anshan, 243000, An'hui China
Corresponding author: Xukun Wu (E-mail: m15755579562@163.com)

ABSTRACT This study proposes preference feedback learning and attribute decoupling control to solve the problems of continuous AIGC video production: fading the historical creative intent, inability to incorporate the user's preference sustainably and content drift due to local edits. On top of multimodal conditional encoding, key-value memory retrieval and latent-space video diffusion generation, we develop a human-computer collaborative video generation system which combines creative intent memory, preference feedback and editable generation control. To model the interaction between different modalities —text, reference images, and historical information—cross-modal attention is used. It adopts short-term caching, long-term memory to maintain creation intention among rounds and adapts short-term caching and temporal constraints with spatial masking for attribute-level local editing and dynamically updates character, action, shot and style preferences from both explicit and implicit feedback. The results of 120 generation tasks indicate that the full model obtained a semantic consistency of 0.341, a 17.4 percentage point higher character retention of 91.8% and an average of 3.5 rounds of interaction compared to single-round prompts. The results show that this model improves cross-round semantic retention and personalized adaptation, and decreases repetitive interaction expenses in the continuous video creation, while maintaining local editing stability.

INDEX TERMS AIGC video creation; creative intent memory; preference feedback; editable generation control; human-computer collaboration

I. INTRODUCTION

Generative artificial intelligence is expanding beyond text and static image generation to video content production, gradually transforming workflows in digital creativity, film and television production, and interactive content design. Bervar, Bertoncel, and Bach (2026), drawing from the context of creative industry research, point out that generative AI has formed a multi-disciplinary research landscape spanning artistic creation, content production, and human-computer collaboration. However, existing research has focused more on application expansion and industrial impact, with insufficient discussion on how user intent is continuously understood and preserved during the process of continuous creation [1]. Rakesh et al. (2025) systematically summarized progress in speaker video generation regarding identity preservation, temporal synchronization, and quality evaluation, while noting that consistency degradation and insufficient controllability still exist under complex dynamic conditions [2]. Oskooei et al. (2025) constructed a scalable generative architecture for multilingual video translation, enhancing cross-lingual video processing capabilities; however, its interactive process remains driven by predetermined task conditions, making it difficult to adapt to the dynamic demands arising from creators' iterative revisions [3]. Naser et al. (2025) further found that different prompting methods affect the content fidelity of generated images and videos, indicating that prompt conditions significantly constrain the generated results; however, relying solely on prompts makes it difficult to consistently maintain semantic coherence across rounds [4]. Oppenlaender, Linder, and Silvennoinen (2025), on the other hand, view prompt engineering as a critical capability in generative creation, emphasizing the impact of prompt design on creative quality; however, this approach remains heavily reliant on users repeatedly organizing and revising instructions [5].

Thus, previous approaches have shown that there are three interrelated research gaps: existing creative intent of historical characters tends to get lost when the number of interaction rounds increases; user preferences are not represented with a continuously updated state; and there is strong

coupling between changes of characters, actions, shots, and style, causing drift towards off-target content when making local edits. To solve these problems, we propose a video creation paradigm for human/computer collaboration, which includes creative intent memory, preference feedback, and generation control that can be edited. To ensure the cross-round semantic consistency, multimodal intent encoding and two-level memory retrieval are used to continuously update individuals' preference from explicit and implicit feedback learning, and characters, actions, shots, and styles are separated as independent editing variables to form a closed loop of generation and dynamic update. Experiments show that this method can achieve semantic consistency and preference alignment, and also achieve the effect of improving local editing stability and interaction efficiency at the same time, giving a feasible modeling path for how to realize AIGC video generation from single-round prompt generation to continuous, controllable, and personalized human-computer collaborative video generation.

II. REQUIREMENTS FOR HUMAN-COMPUTER COLLABORATION AND PARADIGM CONSTRUCTION IN AIGC VIDEO CREATION

When AIGC video creation moves from single-round prompt generation to continuous human-machine collaboration, we need to solve the problems of diminishing historic creativity, the difficulty of continuously integrating individual preferences, and global changes in content due to local changes [6]. To meet these needs, we build a collaborative paradigm that includes three core states, four editing variables, and one closed loop link: firstly, we encode text, reference pictures, and historical modification records into a "creative intention" state, and maintain semantic consistency by jointly searching long term memory and current context; second, the user selection, rating, and editing behavior are mapped to a preference status, which is continually updated as a generating condition; thirdly, the four attributes — characters, actions, shots, and styles — are decoupled into editable control variables. The system works in a cyclic fashion through six phases: "Intention Input — Memory Retrieval — Preference Restriction — Video Generation — Local Edit — Feedback Update," which provides a uniform interface for later Model Calculation and Module Implementation [7].

III. A HUMAN-COMPUTER COLLABORATIVE VIDEO CREATION MODEL INTEGRATING CREATIVE INTENT MEMORY, PREFERENCE FEEDBACK, AND EDITABLE GENERATION CONTROLS

A. MULTIMODAL CREATIVE INTENT ENCODING AND MEMORY

This paper discusses the role of artificial intelligence as a creative tool and a medium for human–machine collaboration from the perspective of artistic practice [8]. In order to solve the problem of inter-cycle intention degradation and inconsistent multimodal representation in continuous AIGC video creation, the model sets up three coding branches for text, reference, and historical interaction, and classifies four kinds of creative attributes — characters, movements, shots, and styles — into explicit semantic slots. Inputs from each modality are mapped to a common feature space through an independent encoder, and feature alignment and fusion are implemented with cross-modal attention. The model employs a two-level storage structure comprising a short-term context cache and a long-term key–value store, which store the current round's dynamic information and cross-round stable intentions, respectively. Their encoding fields and memory organisation are shown in Table 1, providing a unified state representation for subsequent intention retrieval and the construction of generation conditions.

TABLE 1. Multimodal Creative Intent Encoding and Memory Structure

Encoding Object	Input Format	Encoding Unit	Semantic Fields	Memory Structure	Output Representation
Text Intent	Prompts, Modification Instructions	Token Sequence	Characters, Actions, Shots, Style	Short-term memory + long-term memory	Text Intent Vector Z_t
Image Intent	Reference Frames, Character Images, Style Images	Patch sequences	Identity, Scene, Composition, Style	Long-term memory	Visual Intent Vector Z_v
Historical Interactions	History Hints, Selection, and Revision History	Iteration State Sequence	Retain, Modify, Delete	Short-term cache	History State Vector Z_h
Fusion Intent	3-channel encoding results	Cross-modal attention	4 Types of Editable Attributes and Global Semantics	Key-Value Memory Bank	Unified Intent Representation Z_c

B. RETRIEVAL OF CREATIVE INTENT FROM MEMORY AND CONSTRUCTION OF GENERATION CONDITIONS

Prompt Organisation and Retrieval Enhancement Strategies for Multimodal Models Can Improve Information Utilization Under Complex Conditions [9]. Based on the above, the model constructs a joint retriever for two-level storage spaces: a short-term cache and a long-term memory, and generates query vectors using three kinds of features: text, visual and historical state. There are three phases in the search procedure:

(1) Compute the semantic similarity between the query vector and the candidate memory, using the Top-k method to preserve the highly relevant records;

(2) introduce time weights and interactional circular indices to dynamically rearrange the history memory;

(3) Detect and update the priority according to the four semantic slots: characters, actions, shots, and styles.

Then, we combine the filtered memories into five kinds of

generating inputs — global, character, time, shot, and style — and uniformly mapped to the conditional feature space, reserving a special interface for subsequent preference information injection.

C. LEARNING FROM USER PREFERENCE FEEDBACK

The model classifies user feedback into two kinds of explicit signals — rating and selection — and two implicit signals — regeneration and local modification. Preference vectors are constructed around the four attributes: characters, actions, shots, and styles. Preference learning involves three levels of processing:

(1) encoding the four kinds of interaction behaviors into feedback characteristics;

(2) computing attribute preference weights according to acceptance, rejection, and modification directions, and performing incremental updates in combination with historical iterations;

(3) mapping the updated 4-dimensional preference status to the conditional space, in which it co-participates in generation constraints with creative intention [10].

The complete process of feedback collection, feature encoding, weight updating, and condition injection is shown in Figure 1.

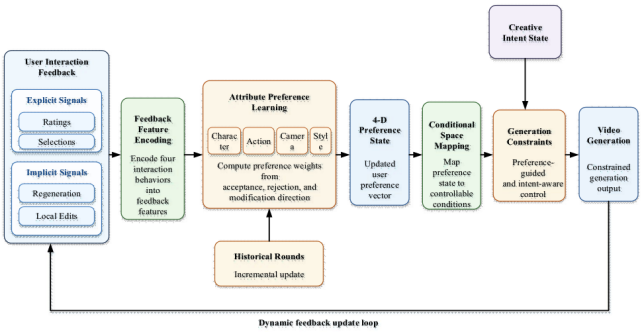

FIGURE 1. User Preference Feedback Learning and Dynamic Update Process

D. AIGC VIDEO GENERATION UNDER PREFERENCE CONSTRAINTS

In order to solve the problem of ensuring that individual preferences always affect spatiotemporal video production, the model adopts a three-layer structure of "Condition Fusion — Hierarchical Injection — Collaborative Constraint." This structure integrates five kinds of generating conditions — global semantics, character identity, action timing, shot parameters, and vision — with four dimensional preference states (character, action, shot, and style) as a common input into the latent-space video diffusion network [11]. There are three phases:

(1) Preference Condition Fusion. The model uses the creative intent vector as the query and the four types of preference representations as keys and values, calculating attribute weights through cross-attention that incorporates feedback confidence:

$$\alpha_i = \frac{\exp\left(\mathbf{q}^{\mathrm{T}} k_i/\sqrt{d} + \beta r_i\right)}{\sum_{j=1}^{4} \exp\left(\mathbf{q}^{\mathrm{T}} k_j/\sqrt{d} + \beta r_j\right)} \tag{1}$$

where α_i represents the attention weight for the ith preference category, q represents the intent query vector, k_i represents the preference key vector, d represents the feature dimension, r_i represents the feedback confidence, and β represents the confidence adjustment coefficient. The model further fuses retrieval conditions, preference states, and historical interaction information:

$$c_p = \mathrm{LN}\left(W_c C + \sum_{i=1}^{4} \alpha_i W_i p_i + W_h h\right) \tag{2}$$

Where c_p represents the preference enhancement condition, C represents the retrieval-generation condition, p_i represents the preference vector for class i, h represents the historical state, W_c、W_i and W_h represent trainable projection matrices, and LN represents layer normalization.

(2) Hierarchical conditional injection. The model establishes three control branches - space, time, and cross - to inject the appearance, movement path, camera motion, and style statistics into the lower, middle, and top layers of the de-noising network, respectively. The noise prediction at step t is expressed as:

$$\begin{aligned}\varepsilon_t = {} & \varepsilon_\theta(x_t, t, c_0) \\ & + s\left[\varepsilon_\theta(x_t, t, c_p) - \varepsilon_\theta(x_t, t, c_0)\right]\end{aligned} \tag{3}$$

where x_t denotes the video latent variable at step t, c_0 represents the spatial condition, ε_θ denotes the spatiotemporal denoising network with parameters θ, and s represents the preference-guided intensity.

(3) The Joint Constraint Optimization Problem. Five loss functions are used to train the model: diffusion reconstruction, character consistency, motion continuity, shot smoothness, and style matching:

$$L = \lambda_d L_d + \lambda_i L_i + \lambda_a L_a + \lambda_m L_m + \lambda_s L_s \tag{4}$$

where L_d、L_i、L_a、L_m and L_s denote the five loss functions, respectively, and λ_d、λ_i、λ_a、λ_m and λ_s denote the corresponding weights. This architecture transforms user feedback into computable, adjustable, and feed-backable generative constraints while preserving independent control channels for the four attributes.

E. EDITABLE GENERATION CONTROLS FOR CHARACTERS, ACTIONS, SHOTS, AND STYLE

The editable generation controls use a separate design for the four attributes — characters, actions, shots, and styles — that allow the user to specify objects for modification while maintaining the stability of non-target content[12-13]. The model establishes character identity parameters P_t, action sequence parameters A_t, shot parameters C_t, and

style parameters S_t, and uses editing intensity coefficients to control the extent to which these four conditions affect the latent features of the current video. The control state after the tth round of editing is represented as:

$$E_t = \alpha_p P_t + \alpha_a A_t + \alpha_c C_t + \alpha_s S_t \tag{5}$$

where E_t represents the comprehensive editing conditions for the tth round; P_t、A_t、C_t and S_t represent the control parameters for characters, actions, shots, and style, respectively; and α_p、α_a、α_c and α_s represent the editing weights for the four attribute categories, respectively. If no change is specified for a specific property, its relative weight will remain 0, thus restricting the involvement of non target properties in the update [14]. In order to resolve the problem of cross-frame drift that frequently occurs when modifying characters and actions, the model further limits the editing magnitude by using the features between adjacent frames:

$$D_t = \frac{1}{N-1}\sum_{i=2}^{N} \left\| (F_i' - F_{i-1}') - (F_i - F_{i-1}) \right\|_2 \tag{6}$$

where D_t represents the temporal variation deviation, N denotes the total number of video frames, and F_i and F_i' represent the features before and after editing for frame i, respectively. Camera control further sets parameters for horizontal displacement x, vertical displacement y, zoom ratio z, and rotation angle θ_4, enabling camera movement changes to be executed independently of character and style conditions. In order to minimize the effect of local modifications on unedited areas of the original video, a preservation loss is adopted during the generation phase:

$$L_{keep} = \frac{1}{N}\sum_{i=1}^{N}(1 - M_i)\left\| F_i' - F_i \right\|_2^2 \tag{7}$$

where L_{keep} represents the preservation loss for non-edited regions, M_i denotes the edited region mask for frame i, $M_i = 1$ corresponds to the target modification region, and $M_i = 0$ corresponds to the preservation region. In this way, the model is able to specify either a single attribute change or a combination of several attributes [15]. Consecutive video frames generated by the substitution, the action, the shot, and the style conversion can be organized as shown in Figure 2 to demonstrate the frame-level change when the four kinds of editing commands are applied to the same video content.

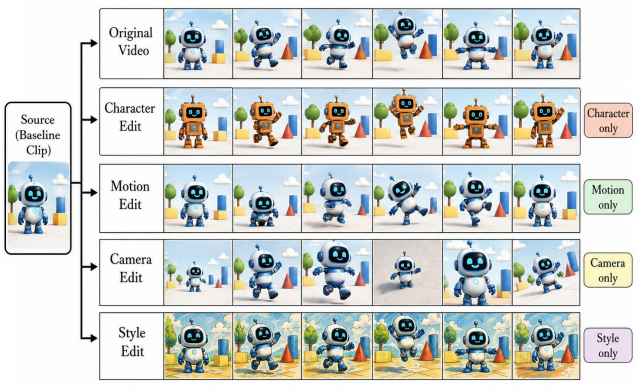

FIGURE 2. Schematic diagram of video frames generated through the editing of characters, actions, shots, and styles

F. CLOSED-LOOP HUMAN-COMPUTER COLLABORATION FOR GENERATION AND DYNAMIC UPDATES

Continuous human-machine collaboration emphasises the co-regulation of creators' decisions and generative systems' responses across multiple rounds of creation; relevant research has discussed the collaborative relationship in which generative artificial intelligence participates in continuous creation from the perspectives of human-machine co-creation and co-agency [16-17]. Accordingly, the human-machine collaboration closed-loop adopts a six-stage cyclical structure comprising 'intent input—memory retrieval—preference constraints—video generation—local editing—feedback write-back'. It rewrites the four types of feedback generated in each round—selection, scoring, re-generation and local modification—back into the state space, thereby transforming the creative process from one-off conditional reasoning to cross-round dynamic computation. After the completion of the tth round, the creative intent state M_t and the user preference state P_t are synchronously updated, with the update rules defined as follows:

$$\begin{aligned} M_{t+1} = {} & (1 - \eta_m) M_t \\ & + \eta_m(\omega_1 I_t + \omega_2 E_t + \omega_3 H_t) \end{aligned} \tag{8}$$

In the equation, M_{t+1} represents the new round of intent memory, I_t represents the current input intent, E_t represents the four types of editing information (characters, actions, shots, and style), H_t represents the historical interaction state, η_m is the memory update rate, and ω_1、ω_2、ω_3 is the fusion weight for the three types of information. The preference state is incrementally adjusted based on two types of feedback: explicit and implicit:

$$P_{t+1} = (1 - \eta_p) P_t + \eta_p \frac{\sum_{j=1}^{K} q_i R_{t,j}}{\sum_{j=1}^{K} q_i} \tag{9}$$

where P_{t+1} denotes the updated preference state, K denotes the number of valid feedbacks in the current round,

$R_{t,j}$ denotes the jth feedback feature, q_i denotes the feedback credibility weight, and η_p denotes the preference update rate. The updated states of the two categories, together with the current editing conditions, jointly construct the input for the next generation round:

$$C_{t+1} = \lambda_m M_{t+1} + \lambda_p P_{t+1} + \lambda_e E_t + \lambda_c C_t \tag{10}$$

where C_{t+1} represents the comprehensive generation conditions for the next round, C_t represents the current conditions, and λ_m、λ_p、λ_e、λ_c control the contributions of the four types of information—memory, preference, editing, and context—respectively. In this way, the system forms a uniform update structure, which includes two kinds of dynamic states, four kinds of editing variables, and a six phase interaction chain, and maintains a state transition trajectory by means of an iterative index. In order to visually represent the transition of the intention – preference states in the feature space, a two-area scatter diagram is used, in which the top area indicates the state distribution before the update, and the bottom area indicates the state distribution after the feedback, as illustrated in Figure 3.

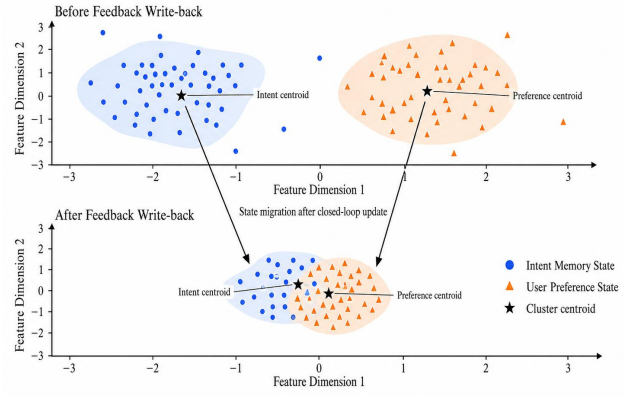

FIGURE 3. Scatter plot comparing the distribution of intention–preference states before and after closed-loop dynamic updates in human–machine collaboration

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. EXPERIMENTAL DESIGN

Five dimensions were studied: generation quality, intent preservation, preference adaptation, local editing and collaboration efficiency. The set up is as follows:

(1) The video creation text data used in the experiment was a self-made AIGC video creation text data set, which included a total of 1, 200 text samples, including 400 video creation prompts, 320 multi-round modification instructions, 240 preference feedback descriptions and 240 partial editing instructions. All the samples were manually labeled in four aspects: characters, actions, shots and style, and were split into training, validation and test sets at the ratio of 8:1:1, which is 960, 120 and 120 samples respectively. There were no existing videos; all experimental videos were created online using the test text.

(2) A total of 120 sets of generation tasks were generated with the 120 creative intentions from the test set, each of which had 3 stages: initial generation, sequential modification and editing of attributes. Compared were five different methods, namely: single-round prompt generation, historical text concatenation, memory-enhanced generation, preference-constrained generation, and full collaborative model. The same number of steps of sampling, random seed configuration and generation backbone were used with each method.

(3) Ablation experiments were conducted in which three core modules (intent memory, preference feedback, and editable controls) were removed, but all other parameters were kept the same as in the full model.

(4) The evaluation metrics comprised seven indicators: text–video semantic consistency, cross-frame temporal consistency, character identity retention rate, edit region retention rate, preference matching rate, average number of interaction rounds, and task completion time.

(5) User Experiment: The experiment involved 24 participants with experience in AIGC video or digital content creation. Each participant completed 30 sets of blind testing tasks, using a 5-point rating scale to record content conformity, controllability of modifications and overall satisfaction, thereby providing a unified evaluation basis for subsequent comparisons of the performance of different methods and individual modules [18].

(6) Experimental Procedures. The four links of the experiment are: data preparation, test generation, special evaluation and user evaluation. After annotating 1,200 text instances across the four attribute categories and dividing the data into training, validation and test sets of 960, 120 and 120 instances respectively, 120 task sets were created with the 120 test intents. Five comparison methods were then run under the same conditions for generation backbone, sampling steps and random seed; subsequently, intent preservation tests were conducted at 1, 3, 5, 7 and 10 rounds, as well as editing intensity tests at five levels (0.2, 0.4, 0.6, 0.8 and 1.0), and three module ablation configurations were completed. Twenty-four people were recruited to complete 30 rounds of blind tests each during the user evaluation stage, and a total of 720 evaluation records were generated. Finally, compile the results of automated evaluation indicators, user scores and task efficiency.

B. COMPARISON OF AIGC VIDEO GENERATION QUALITY ACROSS DIFFERENT METHODS

In order to test the influence of different human-machine interaction on the quality of video, 120 sets of generating tasks were accomplished with the same test text, sampling parameters and random seed conditions. The statistics are presented in Table 2.

TABLE 2. Comparison of AIGC Video Generation Quality Across Different Methods

Method	Semantic Consistency	Temporal Consistency	Character Retention Rate/%	Editing Region Retention Rate (%)	Preference Match Rate (%)
Single-Round Prompt Generation	0.287	0.812	78.4	81.6	72.3
Historical Text Concatenation	0.301	0.826	81.7	83.5	75.8
Memory-Enhanced Generation	0.318	0.847	86.9	86.2	79.4
Preference Constraint Generation	0.326	0.852	87.5	87.1	85.6
Full Collaboration Model	0.341	0.871	91.8	90.4	89.7

Table 2 shows that the full collaborative model achieves the best results in all five metrics. Compared with single round prompt generation, there was an increase of 0.054 and 0.059 in the semantic consistency and in time consistency, and a 13.4, 8.8, and 17.4 percent improvement in the retention rate, edit area retention, and preference matching. Although historical text concatenation can only preserve context to a limited degree, memory enhancement and preference restriction improve the cross-round semantic retention and the personalized condition injection, respectively. This shows that the combination of multi-module constraints can simultaneously reduce content drift and preference bias, thus providing a foundation for further analysis of the intent retention capacity of multiple cycles.

C. MULTI-ROUND RETENTION PERFORMANCE OF CREATIVE INTENT MEMORY

In order to examine the continuing impact of historical creative constraints as the number of interaction rounds increases[19-20]. we conducted 1,3,5,7,10,10 successive revisions on 120 task sets, and calculated the semantic consistency between the current generated results and the original creative intention, as well as the retention rates of the characters, actions, shots, and styles. The performance changes are illustrated in Figure 4.

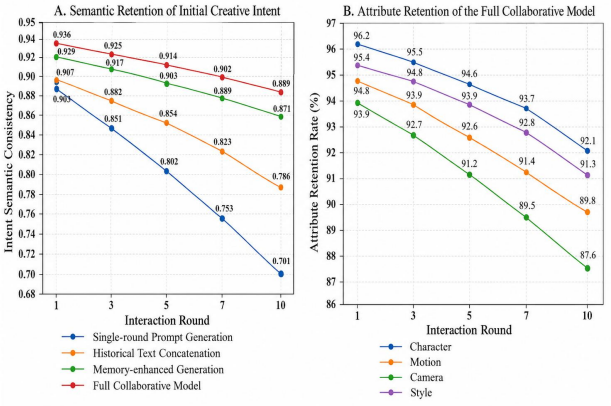

FIGURE 4. Comparison of Multi-Round Retention Performance of Creative Intent Memory

As shown in Figure 4, as the number of interaction rounds increased from 1 to 10, all four approaches showed varying degrees of creative intent decline. On the other hand, the semantic consistency of the whole cooperative model was reduced from 0.936 to 0.889 — a drop of 5.02% — much less than the 22.37% drop seen in the single-round prompt generation. In the attribute level, characters and style retained 92.1% and 91.3% at the 10th round, while the actions and camera shots retained 89.8% and 87.6% respectively. The decrease in camera shot constraints is relatively rapid, suggesting that continuous changes in camera motion are more likely to alter historical spatial relationships, while two-tiered memory retrieval can consistently preserve the identity of the character, the visual style, and the basic action semantics.

D. THE IMPACT OF PREFERENCE FEEDBACK ON PERSONALIZED VIDEO GENERATION

In order to measure the moderating effect of preference feedback on personalized production conditions, we compared four settings — no preference, explicit, implicit, explicit, explicit and implicit — on the basis of 720 sets of blind tests completed by 24 participants. The statistics are presented in Table 3.

TABLE 3. Personalized Video Generation Performance Under Different Preference Feedback Settings

Preference Feedback Settings	Overall Preference Match Rate/%	Character Match Rate (%)	Action Match Rate (%)	Shot Match Rate (%)	Style Match Rate /%	User Satisfaction / 5-point scale
No Preference Feedback	72.3	76.5	71.8	70.2	74.6	3.41
Explicit feedback only	83.1	85.9	82.6	81.7	86.3	4.05
Implicit feedback only	80.4	83.6	79.8	78.9	84.5	3.88
Explicit-Implicit Combined Feedback	89.7	91.2	88.4	87.9	92	4.46

Following the introduction of preference feedback, there was an improvement in personalized matching for all four attributes. In particular, the overall preference matching ratio of explicit and implicit CFT was 89.7%, which is 17.4% higher than the non-feedback scenario. User satisfaction increased from 3.41 to 4.46. The style matching rate was 92.0%, which was higher than that of the actions and shots, which showed that stable visual preferences were more easily obtained from repetitive selection and modification. Only explicit feedback is superior to implicit feedback, which indicates that the rating and selection provide more direct preference supervision, and the combination of the two signals can reduce the bias of preference estimation due to the dependence of a single feedback source.

E. PERFORMANCE ANALYSIS OF EDITABLE GENERATION CONTROL

For 120 tests, we made single attribute edits for four categories — characters, actions, shots, and styles — with a control intensity of 0.2, 0.4, 0.6, 0.8, and 1.0. We calculated the target attribute editing success rate and the non-target content retention rate; Figure 5 shows the results.

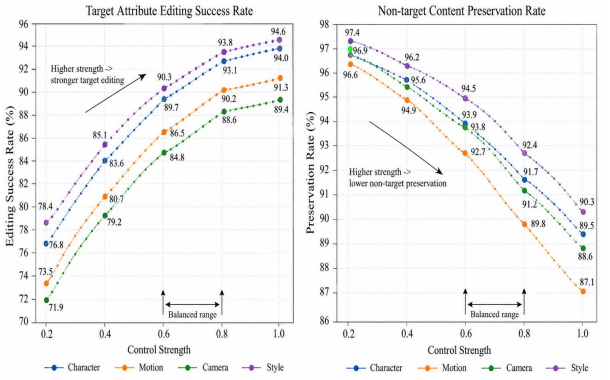

FIGURE 5. Comparison of Editable Generation Control Performance at Different Editing Intensities

As shown in Figure 5, the success rate of editing the four target attributes is continuously increased with the increase of the control intensity from 0.2 to 1.0, with the style and character reaching 94.6% and 94.0%, and the action and shot reaching 91.3% and 89.4% respectively. Moreover, the retention ratio of non target content is reduced to 87.1% under strong control conditions, which shows that the increase of editing weight not only improves the response of target properties, but also increases cross attribute disturbance. At a control intensity of 0.6 to 0.8, editing success rates and content retention rates remain fairly equal for the four categories, providing the basis for an editable control parameter to be used in future collaborative efficiency comparisons.

F. COMPARISON OF CREATIVE EFFICIENCY ACROSS DIFFERENT HUMAN-COMPUTER COLLABORATION PARADIGMS

We calculated the average number of interaction rounds, prompt rewrites, and task completion times for each of the four modes: single, historical, memory, and full cooperation. Then, we used three side-by-side heat maps to demonstrate the efficiency differences in different tasks, as illustrated in Figure 6.

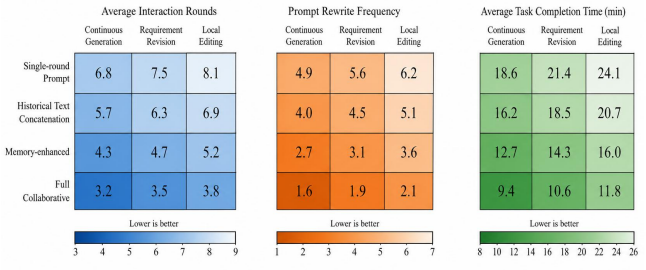

FIGURE 6: Heatmaps of Creative Efficiency Across Different Human-Computer Collaboration Paradigms

Figure 6 shows that the total collaboration model has the least interaction burden. For example, the average number of interaction rounds was reduced from 8.1 to 3.8, and the number of prompt rewrites was reduced from 6.2 to 2.1, from 24.1 minutes to 11.8 mind. Memory enhancements reduced the need to re-enter historic requirements, but additional hints are still needed to correct requests and local editing; full collaboration further combines preference status updates with attribute-level editing, so that the user can directly influence the target conditions. As a result, the operating costs of all three types of task — continuous production, demand modification, and local editing — are significantly lower.

G. MODULE ABLATION EXPERIMENTS

Three ablative configurations were constructed for 120 test tasks — removal of intent memory, elimination of preference feedback, and removal of editable controls — while keeping the rest of the network architecture and reasoning parameters consistent, examining the contribution of each module to the full collaboration model. The results are presented in Table 4.

TABLE 4. Results of Core Module Ablation Experiments

Model Configuration	Semantic Consistency	Preference Matching Rate/%	Editing Success Rate (%)	Non-target Content Retention Rate (%)	Average Number of Interaction Rounds/Round
Intent Removal Memory	0.312	85.8	88.9	88.4	4.8
Removing preference feedback	0.329	79.6	90.2	90.1	4.2
Remove editable controls	0.333	86.7	78.5	80.2	5.1
Full collaboration model	0.341	89.7	91.4	91.3	3.5

In Table 4, the semantic consistency fell from 0.341 to 0.312, and the average number of interaction rounds rose to 4.8, suggesting that cross-round retrieval had a direct impact on the continuous invocation of historical creative constraints; after the removal of the preference feedback, preference matching fell 10.1 percent, which was the biggest drop in this metric among the four configurations; the editing success rate and the non-target content retention rate fell to 78.5% and 80.2% respectively, and it took 5.1 rounds of interaction to complete the target modification. The overall advantage of the whole model was maintained in all four quality measures and interaction rounds, suggesting that the three modules deal with intent retention, personalization constraints, and local modifications; their combination reduces the generation drift and repeated interactions due to the absence of a single module.

H. CASE STUDY OF MULTI-ROUND AIGC VIDEO CREATION

A video creation task was used to conduct a case study with an urban night scene, which comprised eight successive interaction rounds. Following the first generation in Round 1, the character's clothes, gait and zoom speed were changed in Round 2, 3 and 4, and the cold color tone and low-light preference were changed based on the users' choices in Round 5 and 6, and finally characters' positions and composition of the shot were adjusted in Round 7 and 8. The final preference match rate rose from 74.8% to 90.6% while the character identity rate was maintained at 92.7% and the success rate for editing target areas was 92.3% and for non-target areas was 91.1% throughout the 8-round generation process, the semantic consistency of the initial intent moved from 0.934 to 0.905. It took a total of 6 local edits and just 2 prompt rewrites, and lasted 12.9 minutes. This shows that when multiple changes are made on consecutive memory rewritings, memory rewriting, preference updating and attribute control may be combined to decrease the semantic drift and the need for repeated conditional inputs.

V. OPTIMIZATION STRATEGIES FOR HUMAN-AI COLLABORATIVE VIDEO CREATION USING AIGC

Optimizations should be put in place at three levels: prompt conditions, consistency of content, and combinatorial constraints based on results of semantic consistency score of 0.341, intent retention rate of 0.889 over 10 rounds, preference match rate of 89.7%, and shot retention rate of 87.6%.

(1) Standardize the design of high-quality AIGC video prompts, decompose the prompt of natural language instructions into five categories of characters, actions, scenes, shots and styles and set the three-tier template of "fixed attributes—variable attributes—negative constraints". During the multi-round revision process, only the target fields can be modified, and the character identity, core scene, and established style are stored in long-term memory to prevent semantic drift caused by repeated rewriting of the entire prompt.

(2) Increase consistency control of the visual characteristics and content complexity of videos in the same theme. To solve the problem that the shot retention rate drops to 87.6% after 10 interaction rounds, conditional locking of consecutive frames is adopted by using 4 kinds of anchors: character identity features, scene layout features, color tone statistics, and shot trajectories; The intensity of local editing is controlled within 0.6~0.8, and the scope of editing is limited by spatial masks so that the editing of target attributes is synchronous and controllable with non-target parts.

(3) Optimize the generation quality of the specific prompt combination by separately encoding highly coupled preference conditions between different types of attributes (e.g., character–action, character–style, action–shot, shot–scene), and dynamically adjusting the weight of the combination when preference conditions are given, thus allowing prompt words, memory state, and preference feedback to collectively build an updatable generation condition set.

VI. CONCLUSION

A human-computer collaborative video creation paradigm that integrates the memory of creative intent, preference feedback and editable generative control has been developed, enabling the preservation of intent across rounds, the continuous incorporation of individual preferences, and the localised, controllable modification of attributes relating to characters, actions, shots and style. Experimental results indicate that this paradigm demonstrates good overall performance in terms of semantic consistency, preference matching, content preservation and interaction efficiency, highlighting the innovative value of integrating memory states, feedback states and editing variables into a unified

closed-loop generation process. Current experiments remain constrained by the scale of the self-built text dataset, task types and the number of participants; validation regarding complex long videos, cross-scene continuous editing and long-term preference transfer remains insufficient. Future work could further expand to include multiple types of video tasks and real-world creative scenarios, whilst strengthening long-term memory compression, preference drift recognition and multi-attribute coupled editing control, in order to enhance the stability, adaptability and controllability of the continuous creative process.

ACKNOWLEDGMENT

This research was supported by the Humanities and Social Sciences Research Project of Wanjiang University of Technology (WG25064ZD)

REFERENCES

[1] Bervar M, Bertoncel T, Bach P M, "Generative Artificial Intelligence and the Creative Industries: A Bibliometric Review and Research Agenda[J]," *Systems*, vol. 14, no. 2, pp. 138–138, 2026.

[2] Rakesh K V, Mazumdar S, Maity P R, et al., "Advancements in talking head generation: a comprehensive review of techniques, metrics, and challenges[J]," *The Visual Computer*, vol. 42, no. 1, pp. 9–9, 2025.

[3] Oskooei R A, Caglar E, Şahin I, et al., "Generative AI for Video Translation: A Scalable Architecture for Multilingual Video Conferencing[J]," *Applied Sciences*, vol. 15, no. 23, pp. 12691–12691, 2025.

[4] Naser A Y, Sharma S, Niure K, et al., "Geographic prompting and content fidelity in generative Artificial Intelligence: A multi-model study of demographics and imaging equipment in AI-generated videos and images of Canadian medical radiation technologists.[J]," *Journal of medical imaging and radiation sciences*, vol. 56, no. 6, p. 102122, 2025.

[5] Oppenlaender J, Linder R, Silvennoinen J, "Prompting AI Art: An Investigation into the Creative Skill of Prompt Engineering[J]," *International Journal of Human–Computer Interaction*, vol. 41, no. 16, pp. 10207–10229, 2025.

[6] Kranzle T, Sharratt K, "Evaluating Creative Output With Generative Artificial Intelligence: Comparing GPT Models and Human Experts in Idea Evaluation[J]," *Creativity and Innovation Management*, vol. 34, no. 4, pp. 991–1012, 2025.

[7] Chompunuch S, Lubart T, "AI as a Helper: Leveraging Generative AI Tools Across Common Parts of the Creative Process[J]," *Journal of Intelligence*, vol. 13, no. 5, pp. 57–57, 2025.

[8] Avlonitou C, Papadaki E, "AI: An Active and Innovative Tool for Artistic Creation[J]," *Arts*, vol. 14, no. 3, pp. 52–52, 2025.

[9] Son M, Lee S, "Advancing Multimodal Large Language Models: Optimizing Prompt Engineering Strategies for Enhanced Performance[J]," *Applied Sciences*, vol. 15, no. 7, pp. 3992–3992, 2025.

[10] Mzwri K, Szabo T M, "The Impact of Prompt Engineering and a Generative AI-Driven Tool on Autonomous Learning: A Case Study[J]," *Education Sciences*, vol. 15, no. 2, pp. 199–199, 2025.

[11] Vinchon F, Gironnay V, Lubart T, "GenAI Creativity in Narrative Tasks: Exploring New Forms of Creativity[J]," *Journal of Intelligence*, vol. 12, no. 12, pp. 125–125, 2024.

[12] Yan L, Greiff S, Teuber Z, et al., "Promises and challenges of generative artificial intelligence for human learning.[J]," *Nature human behaviour*, vol. 8, no. 10, pp. 1839–1850, 2024.

[13] Luther T, Kimmerle J, Cress U, "Teaming Up with an AI: Exploring Human–AI Collaboration in a Writing Scenario with ChatGPT[J]," *AI*, vol. 5, no. 3, pp. 1357–1376, 2024.

[14] Doshi R A, Hauser P O, "Generative AI enhances individual creativity but reduces the collective diversity of novel content.[J]," *Science advances*, vol. 10, no. 28, p. eadn5290, 2024.

[15] Pitrez C J, "Generative Artificial Intelligence Image Tools among Future Designers: A Usability, User Experience, and Emotional Analysis[J]," *Digital*, vol. 4, no. 2, pp. 316–., 2024.

[16] O'Toole K, Horvát Á E, "Extending human creativity with AI[J]," *Journal of Creativity*, vol. 34, no. 2, pp. 100080–., 2024.

[17] Alfaleh M, "Human–AI Co-Agency for Competency-Based Talent Development: A Framework Integrating Creativity, Cognitive Skills, and Ethical Collaboration[J]," *Journal of Intelligence*, vol. 14, no. 7, p. 144, 2026.

[18] Hikmet Ş, Ozay N, "Human–AI collaboration in architectural design education: towards a conceptual framework[J]," *Buildings*, vol. 16, no. 6, p. 1097, 2026.

[19] Dai C P, Yang M, Lee S, "Designing Human–AI Collaboration for Hybrid Intelligence in Immersive Learning Environments: A Conceptual Framework[J]," *Systems*, vol. 14, no. 6, p. 639, 2026.

[20] McGuire J, De Cremer D, Van de Cruys T, "Establishing the importance of co-creation and self-efficacy in creative collaboration with artificial intelligence[J]," *Scientific Reports*, vol. 14, no. 1, p. 18525, 2024