ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

Intelligent Optimization of AI-Assisted Film and Television Design and Video Production Content

Yi Wang[1] *and Aili Ren*[2,*]

[1] Shanghai Pudong Vocational and Technical College, Shanghai 123456, Shanghai, China
[2] Shanghai Nanhu Vocational and Technical College, Shanghai 123456, Shanghai. China

SUMMARY: *Addressing the practical difficulties which exist in movie and television design and video making-where early idea forming is done quickly, later-stage alterations are very numerous, and information is easy to become wrong between different stages-this paper defines AI-aided content optimization as a cross-stage content problem of decision making instead of a simple comparison of model effects. Based on a review of relevant research on storyboarding, editing semantics, long-form video understanding, controllable generation, and video quality evaluation , we construct a phased evidence corpus tailored to the production workflow and propose a five-dimensional structured coding framework comprising narrative alignment, shot grammar, temporal coherence, perceived quality, and deployment usability.By integrating composite scores, stage-specific adaptation scores, and dimensional gap coefficients, we conduct a unified comparison of the support capabilities of different technical approaches across script design, shot composition, editing and sequencing, quality review, and delivery deployment.The results indicate that current technological strengths are primarily concentrated in front-end visualization and mid-stage continuous content generation, while back-end quality feedback and delivery deployment remain structural weaknesses. Generation and LongVideo are closer to the core of cross-stage collaboration, Story board shows clear superiority in the early period of design, and Quality mainly undertakes back end diagnosis work. This research puts forward that, the future key point of AI-aided movie and TV making ought to move from enlarging single ability to building a closed cycle system that includes quality feedback, lens improvement, and arrangement interfaces.*

KEYWORDS: *AI-assisted film and television design; video production optimization; storyboard pre-visualization; long-form video understanding; quality assessment*

1 Introduction

The working flow for movie and television design and video making is very long. The front portion includes dealing with subject position setting, script splitting, visual style pre-settings, and lens arrangement, hence the back portion needs arranging image materials, holding rhythm control, carrying out quality examinations, and hence transmitting content to platforms. The true difficult point does not only lie in producing a video that can be watched, but lies in guaranteeing that the director's purpose, scene story telling, shot grammar rules, and last product quality keep consistent in many production repeated processes. In real-world projects, planners, directors, art directors, editors, and post-production teams typically work in separate phases. As the consequence, the same content is undergone multiple rewritings between the

*tiancai1027@hotmail.com
https://doi.org/10.65102/is2026754
Published: 30 April, 2026

script, the storyboard, the rough cut and the final cut, therefore it leads to a continuous accumulation of information loss and semantic drift. This problem is especially obvious in situations like the industrialized manufacturing of short videos, movie and television pre-visualization, educational video planning, and brand content making, where project time lines are pressed and the expense of manual text checking increases, hence making the problem of fast pre-production concept forming followed by large post-production changes even more sharp. Therefore, AI is more and more being put into links such as role design, picture draft of frames, reference picture making, frame pre-display and video synthesizing, with the goal of at the same time raising pre-production efficiency and mid-production producing speed [1-5].

In response to this need, current research has established a relatively clear technical trajectory centered on cinematic language and editing semantics.Related work attempts to transform shot size, camera position, shot transitions, and transition relationships from empirical rules into computable entities, thereby supporting footage organization, automatic editing, and assisted sequence assembly [6-8]. The significant value of these methods lies in the fact that they no longer view video clips as mere visual streams but rather treat shot organization as an expressive system with a grammatical structure.For film and television design, this means that the translation from script to shot can partially move beyond reliance on pure experience; for video production, it means that automated systems have the opportunity to participate in shot selection, sequence ordering, and rough cut generation. However, most existing methods remain at the level of shot relationship modeling and clip-level decision-making, failing to sufficiently incorporate upstream narrative objectives or provide adequate downstream quality feedback, and thus have not yet developed true cross-stage collaborative capabilities.

At the same time, studies on long-time video comprehension and movie-level story analysis are continuously going forward, hence researches have discussed problems including piece-level cutting, time point positioning, cross-piece memory keeping, and long-distance dependence model building. [9-14]. Such progress is critical for film and television content production, as the quality of the final product depends not only on the visual completeness of individual shots but also on the sustained consistency of characters, actions, scenes, and events across linear time.Long-form video models provide a new technical foundation in this regard: they can more reliably preserve plot threads, scene evolution, and cross-segment information relationships, while also offering intermediate representations better suited for the production pipeline in video content retrieval, plot summarization, and shot relocalization. However, this line of research places greater emphasis on semantic retention and retrieval efficiency in the temporal dimension, with relatively limited attention paid to shot syntax, subjective quality perception, and post-production editing interfaces.As a result, while systems may be able to determine whether content is coherent, they may not necessarily be able to assess whether a shot is valid, whether footage is usable, or whether the output is suitable for direct entry into rough-cut or fine-tuning workflows.

On the production input side, studies about integrated video making and revising, multi-camera video generating, movement route controlling, and long-length content generating have further pushed AI from single-camera demonstrations to continuous content making [15–18]. When we compare with early methods which could only synthesize short clips or static-style segments, these current methods have displayed relatively big enhancements in character consistency, shot transitions, action continuity, and conditional control. As for film and television design and video making work, this tells us AI has already not been only a tool for drawing conceptual drafts, but it starts to show the capability that it can finish storyboard visual making, plot part early watching and fast first cutting producing. Therefore, the research focus has had a change: whether the produced output can correctly reflect the script, whether the

connections between shots match story rhythm, whether the images satisfy aesthetic and technical quality rules, and whether the outcomes can be stably reused in current editing software, checking work processes, and delivery situations have slowly become more urgent problems than the simple question of "whether generation can be done"[19-24]. If these problems are still not solved, the promotion of model abilities will have difficulty in being changed into synchronous promotion of production efficiency and final video quality.

The deficiencies of current researches mainly concentrate on three interconnection levels. Firstly, the pre-production design, long-form video comprehension, video generation, video editing, and quality evaluation still are comparatively separated technical modules. The absence of common objective goals and unified judgment standards in different phases has caused the fast growth of front-end methods, but back-end check and transmission support still stay comparatively weak. Second, although time consistency has already become a core goal for the majority of systems, the shot syntax and the feeling-based quality have not been combined into one united optimization frame hence. Therefore, a lot of outputs can be coherent in narrative terms, but they still are not connected in audiovisual expression, hence they cannot directly satisfy the overall demands of film and television making-that is, "can be narrated, can be watched, can be delivered." Third, although current video quality evaluation approaches can find technical distortions, text alignment errors, and subjective perception defects, they seldom send these diagnosis results directly back to shooting adjustments, revision reminders, and editing decisions. Multi-agent nonlinear edition has already started to fill this gap, but there still exists a quite big space before we can get a stable production feedback loop [25].

Based on the above assessment, this paper frames "AI-assisted intelligent optimization of film and television design and video production content" as a cross-stage content decision-making problem.The focus of this paper is not on further expanding the generative capabilities of any single model, but rather on how to simultaneously characterize narrative alignment, shot grammar, temporal coherence, perceived quality, and deployment usability within a single analytical framework, and how to map these capabilities to production stages such as script design, storyboarding, editing, quality review, and delivery deployment.To achieve this goal, this paper organizes a highly relevant corpus covering storyboarding, editing semantics, long-form video understanding, controllable generation, and quality evaluation. Based on this, we propose a five-dimensional structured encoding and stage-based scoring mechanism to identify the strengths and weaknesses of current technological capabilities, as well as the production nodes most suitable for prioritized intervention.This work focuses on three main aspects: first, constructing a stage-based evidence corpus and object organization method tailored to film and television design and video production; second, proposing a cross-stage multi-objective content optimization model that translates heterogeneous technical approaches into unified production-stage capabilities; and third, refining intelligent optimization priorities and deployment directions applicable to actual production chains based on category comparison, gap diagnosis, and stage weighting.

2 Research on AI-Assisted Intelligent Optimization of Film and Television Design and Video Production Content

2.1 Definition of Research Object and Organization of Evidence Corpus

After artificial intelligence is merged into film and television design and video making, one group of operable abilities has appeared on many different production nodes. Corpus choosing starts from real-world making chains, regarding script splitting, storyboard pre-seeing, shot arranging, material gathering, long-video comprehension, last video making, quality checking,

and sending arrangement as continuous steps inside one problem chain. Therefore, the research topics are divided by us into five kinds: firstly, pre-production design instruments that from text summaries to storyboards; second, the methods that are used to comprehend the language of cinematography and the semantics of editing; thirdly, the models that are used for comprehending long videos of cinema size and the memory of time; Fourth, the controllable multi-time generation and unified video making framework; and fifthly, the quality examination and subjective assessment working procedures for AIGC videos. The aim of this method is to guarantee that the analysis which comes after directly corresponds to particular tasks in real production, instead of being still restricted within a broad gathering of "methods which are related to video".

During the screening of candidate literature, this paper first excluded three categories of materials. The first category consists of entries with weak relevance to film and television scenarios that merely demonstrate general generative capabilities; while such research can illustrate the overall progress of video models, it cannot directly support shot composition, editing decisions, or final film review.The second category consists of studies that report only example results without clear definitions of inputs and outputs or evaluation metrics, as such research is difficult to incorporate into a unified coding scheme. The third category includes works that overlap with the core problem of this paper but primarily serve compression and transmission, general retrieval, or low-level visual representation. While these methods have technical value, they cannot directly address the issue of cross-stage content optimization in film and television design and video production.After initial screening and review, 25 highly relevant references were retained. Of these, 20 were included in the core coding, while 5 served as background reviews, general benchmarks, or protocol constraints and did not directly contribute to quantitative scoring. The core coding items covered storyboarding [2-5], editing semantics [6-8], long-form video understanding [11-14], controllable generation [15-18], and quality evaluation [19-23]; supporting items primarily include reviews, cinematic-level benchmarks, and subjective evaluation protocols [1, 9, 10, 20, 24]. Sources of evidence corpora and organization of objects. As shown in Table 1.

Table 1: Sources of Evidence Datasets and Organization of Objects

Category	Literature Reference	Quantity	Direct Data Objects	Role in This Paper
Film Design and Storyboarding	[1]-[5]	5	synopsis, storyboard keyframe, camera preset, interaction log	Explain the collaborative needs of pre-design tasks and storyboarding
Shot Language and Editing Semantics	[6]-[8], [25]	4	shot tag, transition relation, auto-edit decision	Provide the syntax of shots, assembly relations, and editing closure basis
Long Video Understanding and Narrative Consistency	[9]-[14]	6	movie hierarchy, long-video QA, temporal grounding, stream memory	Provide basis for scene hierarchy, temporal positioning, and narrative continuity
Controllable Generation and Multi-Shot Composition	[15]-[18]	4	unified condition, shot token, motion trajectory, storyboard-conditioned generation	Provide basis for content generation and controllable correction
Video Quality and Subjective Evaluation	[19]-[24]	6	AIGC-VQA dataset, LMM benchmark, technical VQA, subjective protocol	Provide basis for perceptual quality, technical quality, and evaluation protocols

Note: To avoid an amplification effect on intensity scores caused by reviews, protocols, and general benchmarks, [1], [9], [10], [20], and [24] are not included in the quantitative coding of the 20 core pieces of evidence; they serve only as background, benchmarks, and protocol constraints.

To bring heterogeneous technologies into the same analytical framework, this paper organizes the core corpus into a three-tier structure, as shown in Figure 1.

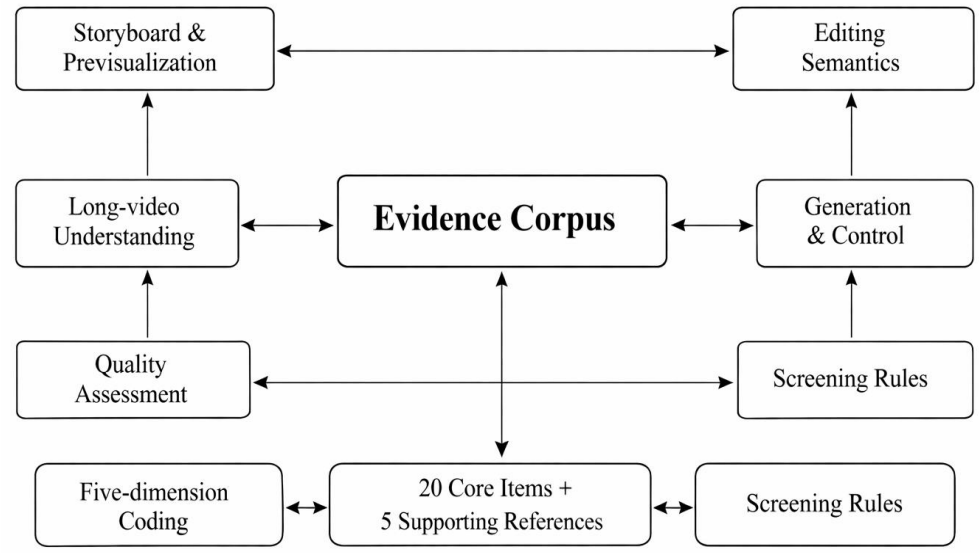

Figure 1: Mechanism for constructing the evidence corpus of AI-assisted film and television design and video production.

The first layer is the technology category layer, which classifies the 20 core pieces of evidence into five categories: Storyboard, Editing, LongVideo, Generation, and Quality. This classification is not based on the conference to which the paper belongs or its publication year, but rather on the production phase it directly serves and its output object.For example, entries in the Storyboard category output shot sketches, keyframes, camera position references, or director collaboration interfaces; entries in the Editing category output shot labels, transition relationships, automatic splicing decisions, or nonlinear editing suggestions; entries in the LongVideo category output scene hierarchies, long-term memory, temporal localization, and streaming understanding results;Generation-class entries produce controllable shot segments, multi-shot videos, or results generated under motion trajectory constraints; Quality-class entries produce perceived quality scores, video-text consistency assessments, technical quality diagnostics, or subjective evaluation criteria. Organizing the corpus by object rather than by algorithm name helps avoid the issue where different technical approaches appear nominally similar but have entirely different actual objectives when compared.

The second layer is the optimization dimension layer. This paper maps each piece of core evidence to five dimensions: narrative alignment, shot syntax, temporal coherence, perceived quality, and deployment usability. Narrative alignment is used to characterize the degree of consistency between the script, scene intent, and generated content; shot syntax is used to characterize the computability of shot types, camera angles, transitions, editing rhythm, and visual organization;temporal coherence describes the degree of temporal continuity among characters, actions, audiovisual events, and scene transitions; perceptual quality describes clarity, visual harmony, textual consistency, and subjective viewing experience; and deployment usability describes real-time performance, interactivity, human-machine collaboration, software workflow compatibility, and deliverability.Together, these five dimensions cover the core requirements of film and television production, from content development to process implementation. Retaining only the first four dimensions risks overestimating the model's performance in a laboratory setting; conversely, emphasizing deployment usability alone may obscure narrative and cinematographic shortcomings in the

content itself. Therefore, the five-dimensional framework serves the research objective of simultaneously achieving the three levels of "capability, quality, and implementation."

The third stratification is the production stage. For making match with real world production work flows, this article further carries on mapping of dimension score values to five concrete stages: script design, shot arrangement, cut and arrangement, quality examination, and send out and application. The design of script puts emphasis on narrative accordance and explicit content scope; The composition of pictures lays stress on the syntax of cinematography and the arrangement of visual elements; Editing and arranging stress the logical links and time matching between each shot; quality examination puts stress on people's feeling of quality and the finding of flaws; and the delivering and the deploying emphasize the system's stable calling, interactive feedback, and output reusing inside the real producing flow. After this three-layer structure has been set up, single research papers are not any longer only "methods in a special direction," but they are changed into functional units which have the ability to be put together into production work flows for comparison.

For making scoring have consistent standard, this paper uses a three-layer grade system which is made of 0, 1, and 2. A numerical value of 0 shows this dimension is not able to give a direct contribution; a mark of 1 shows that the dimension is got involved in a supporting ability or functions only inside special modules; and a numerical value of 2 shows that this dimension forms the main optimization target of our research. Taking a shot-generation system as an example, if its core task is to map a textual synopsis to shot sketches while explicitly controlling shot composition and camera movement, then narrative alignment and shot syntax can be assigned a score of 2; if it merely supplements the results with simple temporal connections, then temporal coherence is typically assigned a score of 1.Similarly, for a quality evaluation model, if its primary outputs are video quality scores and quality diagnostics, Perceived Quality is assigned a value of 2; if Deployment Usability is reflected only in reported inference efficiency or human-computer preference surveys, it is assigned a value of 1. By adopting this rule, we can distinguish between "directly undertaking a specific task" and "incidentally involving a specific capability," thereby reducing the interference of generalized descriptions on subsequent comparisons.The five-dimensional coding rules and comprehensive weight configuration are shown in Table 2.

Table 2: Five-Dimensional Coding Rules and Comprehensive Weight Configuration

Dimension	Meaning	0 Points	1 Point	2 Points	Comprehensive Weight
Narrative Alignment	Consistency between script, storyboard, and video semantics	Not addressed	Local text or scene constraints	Explicitly modeling as a core goal	0.25
Shot Grammar	Rules for shot size, camera position, transitions, and shot organization	Not addressed	Local shot control or tag usage	Explicitly encoding shot grammar or editing semantics	0.20
Temporal Coherence	Temporal positioning, action continuity, and long-range consistency	Not addressed	Local temporal modeling	Core temporal/memory mechanism	0.25
Perceptual Quality	Technical quality, aesthetic quality, and subjective preference	Not addressed	Occurs as an auxiliary means	Constitutes the core evaluation goal	0.20
Deployment Readiness	Real-time capability, interactivity, human-machine collaboration, and deployment closure	Not addressed	Reporting efficiency or user research	Explicitly aimed at delivery or interaction design	0.10

2.2 Building a cross-phase intelligent optimization model

After organizing the objects, the problem is transformed into: how to express the local capabilities of different categories of technologies as a unified score that allows for cross-comparison. Simply listing research progress by category does not allow us to determine to what extent a particular technology simultaneously supports narrative, cinematography, timing, quality, and implementation. To this end, this paper first defines a comprehensive optimization score for each individual piece of core evidence to characterize its cross-dimensional support strength. Its definition is shown in Equation (1).

$$S_i = w_n n_i + w_c c_i + w_t t_i + w_p p_i + w_r r_i \tag{1a}$$

$$w_n + w_c + w_t + w_p + w_r = 1 \tag{1b}$$

In the equation, S_i represents the comprehensive optimization score of the i th core piece of evidence; n_i, c_i, t_i, p_i, and r_i respectively denote the encoded values of this evidence in narrative alignment, shot grammar, temporal coherence, perceptual quality, and deployment usability, with a range of {0, 1, 2} for each; w_n, w_c, w_t, w_p, and w_r are the weights for the five dimensions.Considering the characteristics of film and television design and video production tasks, this paper sets the five weights to 0.25, 0.20, 0.25, 0.20, and 0.10. This configuration is based on two considerations. First, narrative alignment and temporal coherence determine whether the content holds together, so they are assigned relatively higher weights;Second, while cinematographic syntax and perceived quality are equally critical, the former primarily determines the mode of expression, while the latter primarily determines the viewing experience; thus, they are assigned the second-highest weights. Deployment usability is retained separately but not given an excessively high weight to avoid engineering constraints unduly influencing content evaluation. The purpose of Equation (1) is not to simply label the method as "high" or "low," but to provide a unified metric for mapping subsequent stages.

A composite score alone is insufficient to reflect the differences between stages in the production workflow. The focal points of film and television design and video production vary across different phases; a single technology may be highly valuable in script design but may not be applicable during delivery and deployment. Therefore, this paper further defines a stage-specific suitability score to characterize the overall level of support provided by the 20 core criteria for a given production stage, as shown in Equation (2).

$$M_s = \frac{1}{m}\sum_{i=1}^{m}\left(a_{s,n} n_i + a_{s,c} c_i + a_{s,t} t_i + a_{s,p} p_i + a_{s,r} r_i\right) \tag{2}$$

In the formula, M_s represents the average stage adaptation score for the s th production stage; m represents the number of core evidence items, set here to 20; $a_{s,n}$, $a_{s,c}$, $a_{s,t}$, $a_{s,p}$, and $a_{s,r}$ represent the weight requirements for the five dimensions in stage s. The five stages defined in this paper are script design, shot composition, editing and sequencing, quality review, and delivery and deployment. Correspondingly, the phase weights for script design are set to (0.35, 0.25, 0.15, 0.10, 0.15), highlighting the foundational role of narrative alignment and shot organization in the early stages; shot composition is set to (0.20, 0.30, 0.25, 0.10, 0.15), emphasizing shot size, camera angles, and temporal relationships;Editing and sequencing is set to (0.15, 0.35, 0.20, 0.10, 0.20), allowing both shot syntax and deployment interfaces to contribute to the score; Quality review is set to (0.10, 0.10, 0.20, 0.45, 0.15), highlighting the dominant role of perceived quality in the back-end review;Delivery and deployment are set to

(0.10, 0.05, 0.20, 0.20, 0.45), highlighting system stability and reusability. The introduction of Equation (2) allows the same set of evidence to be re-projected onto different stages, thereby identifying "which type of technology is more suitable for which segment of the production chain."

Beyond stage scores, this paper also needs to identify genuine structural weaknesses, as a high average does not imply the absence of significant gaps. To prevent strong dimensions from masking weak ones, this paper further defines a dimension gap coefficient, as shown in Equation (3).

$$G_k = 1 - \frac{1}{2m}\sum_{i=1}^{m} x_{i,k} \tag{3}$$

In the formula, G_k represents the gap coefficient for the k th optimization dimension; $x_{i,k}$ represents the encoded value of the i th core piece of evidence on the k th dimension; and the 2 in the denominator represents the theoretical maximum score for a single dimension. The closer G_k is to 1, the more insufficient the supply of that dimension is in the current technological ecosystem; the closer it is to 0, the more fully covered that dimension is.

The key enhancement of that model is located in changing different technical proofs into comparable things under a united production logic. Current researches, therefore, always carry out in task-special domains: the frame-by-frame researches entirely concentrate on pre-rendering, the long-form video researches entirely place emphasis on temporal modeling, and the quality estimation works entirely focus on diagnostic accuracy. This therefore leads to an absence of comparable common coordinate systems across varied research directions. But the method that this paper puts forward, therefore, puts five kinds of technologies into a one dimension space, and after that it uses stage mapping to make the dimension scores map back to their position relations which are inside the production working flow.

In the process of carrying out this work, this paper moreover has put restrictions on two factors which are able to influence the stability of research conclusions. First, all coding works were carried out on the basis of this paper's main tasks, core modules, and primary evaluation standards, hence they were not based on the generalized descriptions which are in the abstract. Second, although stage weights have the function to serve production practices, they are not directly given by individual cases, but on the basis of the primary goals of each stage in film and television production, they are set out. This method decreases the influence of subjective inclinations upon the outcomes and supplies clear, can-be-traced objectives for the afterward robustness verification work. The cross-stage intelligent optimization frame, therefore, finally puts scripts, storyboards, shot semantics, generated outcomes, quality feedback, and delivery restrictions into one united analytical mechanism, hence outputting directly explainable adjustment results on the stage level. The cross-stage multi-goal content intellectual optimization frame is shown in Figure 2.

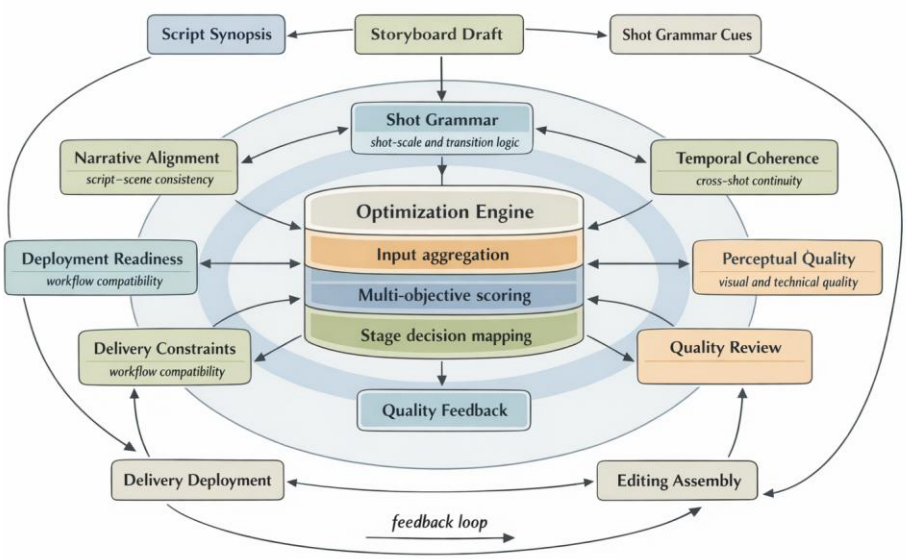

Figure 2: Cross-stage multi-objective intelligent content optimization framework.

2.3 Evaluation Protocol, Robustness Verification, and Result Output Design

For guaranteeing the aforementioned scoring outcomes can be directly put into follow-up comparisons, this research splits the evaluation flow into item-level, category-level and stage-level constituent parts. On the single item level, five-dimension scoring is carried out for every one of the 20 core measurement indicators; on the category level, dimension mean values are computed for the five technique categories-Storyboard, Editing, LongVideo, Generation, and Quality-and the overall optimization score is got through the employment of Equation (1): S_i ;At the stage level, the category averages are substituted into Equation (2) to calculate the adaptation scoresM_s for the five stages: script design, shot composition, editing and splicing, quality review, and delivery and deployment. As shown in Figure 3, the evaluation process consists of four parts: item scoring, category aggregation, stage mapping, and result output.

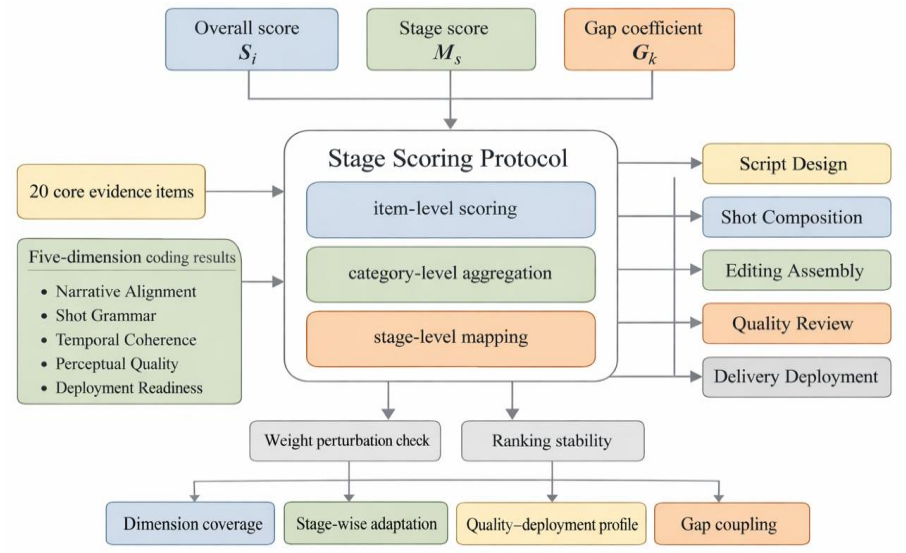

Figure 3: Phase Scoring Protocol and Result Flowchart.

Stage result points all become converted to percentage scales, in which 0 point, 1 point, and 2 points separately correspond to 0%, 50%, and 100%. This transformation does not change the original ordering, but it functions to promote reading convenience among various different categories. Concurrently, this paper calculates the dimensional gap coefficient G_k based on Equation (3) to identify capability items that consistently score low across multiple technical categories. Subsequent result analysis will combine M_s and G_k to determine the primary sources of stage-level weaknesses.

The verification of robustness is carried out through the utilization of weight perturbation. When we maintain the overall weight to be unchanged, the weight of deployment availability is adjusted from 0.10 to 0.20, hence the weights of narrative alignment and perceived quality are each decreased by 0.05. The values of S_i and M_s are then recalculated. If the category ranking and phase priorities remain fundamentally unchanged, the aforementioned conclusions are considered robust.

3 Analysis of Cross-Phase Optimization Results and Production Adaptation

3.1 Analysis of Technical Capability Distribution and Overall Optimization Focus

As what Table 3 displays, the average scores of the five technique sorts on the five optimization aspects are not equal, and the total combination scores also show obvious layer division. The generation item obtains a composite score which is 1.29, it ranks the first; Storyboard and LongVideo have obtained scores of 1.09 and 1.05, respectively, hence they are placed in the middle position; Editing and Quality both get the score of 0.89, therefore this shows the overall support ability is relatively lower. This result gives us the enlightenment that the technology which at present is nearest to the center of cross-stage content optimization is not a single-method method that only focuses on shot editing or quality diagnosis, but it is a generative method that can at the same time deal with narrative constraints, continuous content organization, and visual output.

Table 3: Dimension Means and Comprehensive Optimization Scores for the Five Technology Categories

genre	Narrative	ShotGrammar	Temporal	Perceptual	Deployment	Overall (O)
Storyboard	1.75	1.75	0.50	0.50	0.75	1.09
Editing	0.50	2.00	0.75	0.50	0.75	0.89
LongVideo	1.50	0.25	2.00	0.00	1.25	1.05
Generation	1.25	1.25	2.00	1.00	0.25	1.29
Quality	0.75	0.00	1.00	2.00	0.50	0.89

When we observe the average numerical values on each dimension, the strong points of each category are very obvious. Storyboard method obtains 1.75 scores on both Narrative Alignment and Shot Grammar, therefore this indicates that this kind of method is more suitable for front-end work tasks including script decomposition, storyboard generation, and shot previews. Editing has obtained a score of 2.00 on Shot Grammar-the maximum among all groups-this shows that shot marking, transition connections, and automatic editing rules still mainly lie in the scope of editing-semantic methods. LongVideo obtains a mark of 2.00 on Temporal Coherence and 1.25 on Deployment Readiness, therefore this shows its long-term

memory, cross-segment position arrangement, and continuous processing abilities are more in accordance with the demands of continuous work in actual production work flows. The generation work also got a score of 2.00 on the item of Temporal Coherence, hence it got a score of 1.25 on both Narrative Alignment and Shot Grammar, therefore it shows that it has the ability to deal with both content continuity and shot organization. In terms of Quality, Perceptual Quality has attained 2.00, yet Shot Grammar has achieved 0.00, this shows that this kind of method is more suitable for judging whether results can be used, instead of directly carrying out shot-level organization works. The distributing condition of encoding of the core evidence on the five optimization dimensions is displayed in Figure 4.

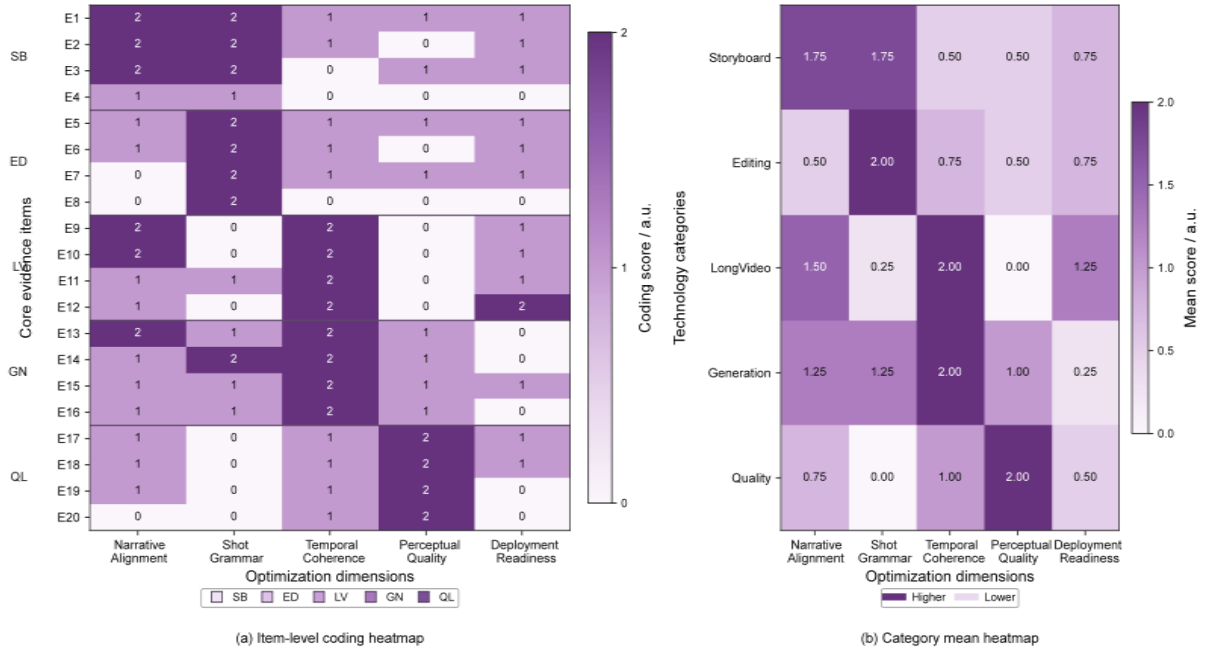

Figure 4: Encoding distribution of core evidence across the five optimization dimensions.

In the Figure 4(a), the valid covering ratios for Narrative Alignment and Temporal Coherence both attain 85.0%, therefore Shot Grammar is 65.0%, thus Perceptual Quality is 60.0%, and hence Deployment Readiness is 50.0%. This distribution shows that the nowadays research has obtained big progress in the fields of narrative continuity and time coherence, but the quality feedback and sending interfaces-which are more near to the backend work flow-are still comparatively weak. The category average heat displayed in Figure 4(b) further enlarges this difference: high numerical values for Storyboard are mainly gathered in the Narrative Alignment and Shot Grammar columns; The high numerical values of LongVideo and Generation are concentrated together in the column of Temporal Coherence; and high numerical points for Quality are gathered in the Perceptual Quality column. The dissimilar technical groupings display divided gathering instead of even spreading in the five-dimension space. This point means that cross-stage optimization at present depends on the cooperation of multi-modal methods, not a single technical method, to realize the coverage of the whole process.

The clusters of high values in Figure 4 also reveal the current focus of research. The high values in the front end are concentrated in the "Storyboard" direction, indicating that the translation from text to shots in film and television design has reached a high level of maturity; the high values in the middle section are concentrated in the "Generation" and "LongVideo" directions, indicating that continuous content generation and long-sequence understanding have become the areas advancing most rapidly; and the high values in the back end are concentrated

in the "Quality" direction, indicating that video quality assessment has gradually evolved into an independent research branch.This also exposes a problem: the high-value areas in the front-end, middle-end, and back-end are dominated by different categories, and these categories have not yet naturally converged. This leads to a disconnect in the production chain, where "front-end solutions are developed quickly, middle-end video can be generated continuously, and the back-end can perform scoring, but there is a lack of stable feedback between the three."Table 3 and Figure 4 collectively demonstrate that the current technological ecosystem possesses the capability elements required for cross-stage optimization, but these elements remain scattered across different technology clusters and have not yet formed a unified production organization logic.

3.2 Analysis of Adaptation Differences and Key Gaps Across Production Stages

When category averages are further projected onto the production workflow, stage-specific differences reflect production adaptability more directly than dimensional averages. The stage adaptation curves and relative deviations for different technical approaches are shown in Figure 5.

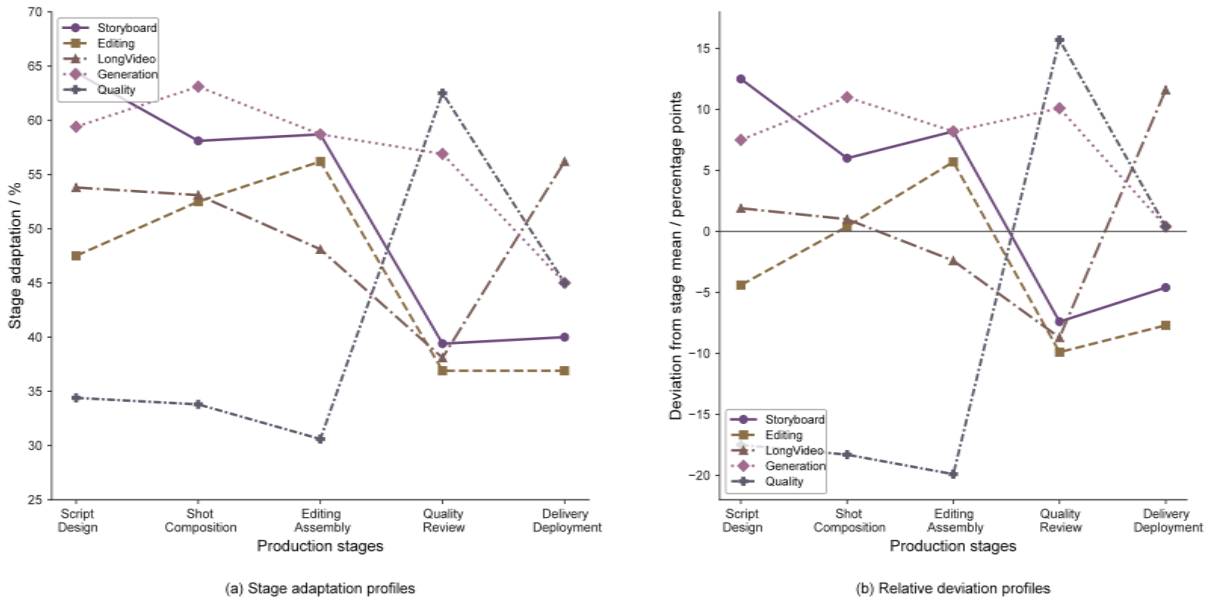

Figure 5: Stage adaptation curves and relative deviations for the five technology categories.

In Figure 5(a), the overall average adaptation scores for the five stages are, in order: Script Design 51.9%, Shot Composition 52.1%, Editing Assembly 50.5%, Quality Review 46.8%, and Delivery Deployment 44.6%.The average alignment scores for the front and middle stages are significantly higher than those for the back end, indicating that as AI is integrated into film and television design and video production, the first areas to benefit are content organization and shot composition, rather than back-end review and delivery.

The strengths of the five technologies vary across stages. Storyboard achieved 64.4% in Script Design, 58.1% in Shot Composition, and 58.7% in Editing Assembly, indicating that this method is best suited for deployment in the front end to rapidly define content boundaries and shot layouts.Generation achieves 63.1% in Shot Composition, 59.4% in Script Design, and 58.7% in Editing Assembly, indicating that multi-shot generation and unified creative frameworks are no longer merely about "producing footage," but can build upon the initial structure and directly

support mid-stage assembly.Editing achieved 56.2% in Editing Assembly, higher than in other stages, consistent with its role in shot syntax and sequence rearrangement. LongVideo reached 56.2% in Delivery Deployment, the highest value in this category, indicating that long-sequence understanding and continuous reasoning mechanisms better align with production environment requirements for stable processing and context maintenance.Quality achieved 62.5% in Quality Review, significantly higher than the other four stages, indicating that its most suitable role remains the backend review interface rather than frontend conceptualization or mid-stage generation.

The relative deviations in Figure 5(b) further illustrate the boundaries of category division. Quality scored 17.5, 18.3, and 19.9 percentage points below the stage averages in Script Design, Shot Composition, and Editing Assembly, respectively, yet exceeded the stage average by 15.7 percentage points in Quality Review, indicating that its strengths are highly concentrated in the backend.LongVideo outperforms the stage average by 11.6 percentage points in Delivery Deployment but underperforms by 8.7 percentage points in Quality Review, suggesting it functions more as a module for continuous understanding and maintenance within the production chain.Storyboard outperforms the stage average by 12.5 percentage points in Script Design, but falls short by 7.4 and 4.6 percentage points in Quality Review and Delivery Deployment, respectively, indicating strong early-stage design capabilities but insufficient extension into the back end.Generation outperforms the average in the first three stages, with Shot Composition exceeding the average by 11.0 percentage points, indicating that it is best suited to serve as the mid-stage hub connecting storyboards to visual clips.

The stage curves and deviation charts together provide a clear structural assessment: the current technological ecosystem does not feature a scenario where "all stages are dominated by a single category." The front end is dominated by Storyboard, the middle stage is jointly supported by Generation and Editing, and the back end sees Quality and LongVideo playing key roles in different capacities.This division of labor does not imply that system design must be fragmented; on the contrary, it suggests that cross-stage optimization is better suited to organizing capability modules according to the responsibilities of each stage in the production chain. If one were to forcibly compress front-end conceptualization, mid-stage generation, and back-end review and deployment into a single model, the most likely outcome would not be overall enhancement, but rather the loss of competitive advantage as capabilities across certain stages become averaged out.

3.3 Analysis of Adaptation Differences and Critical Gaps Across Production Stages

Areas where both the dimensional mean and stage scores are low constitute the parts of the current production chain that most urgently require reinforcement. The Perceived Quality-Deployment Availability Status Map and the Narrative-Temporal Support Curve are shown in Figure 6.

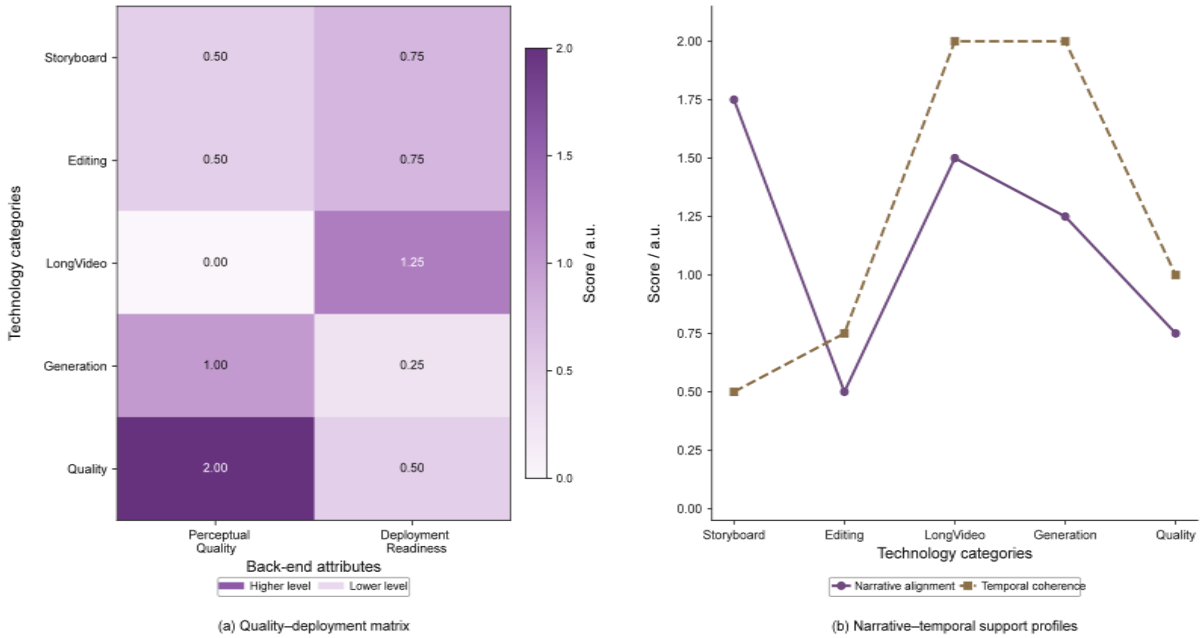

Figure 6: Perceived Quality-Deployment Availability Status Map and Narrative-Temporal Support Curve.

In Figure 6(a), Quality scores 2.00 for perceived quality-the highest among the five technology categories-but only 0.50 for deployment availability; LongVideo scores 1.25 for deployment availability-the highest across all categories-while its perceived quality is 0.00; Generation has a perceived quality of 1.00, but its deployment availability is only 0.25; Storyboard and Editing rank in the lower-middle range on both dimensions. This distribution indicates that backend quality, engineering implementation, and continuous content support are not concentrated within a single technology cluster but are instead handled by different categories. The Narrative-Temporal Support Index shown in Figure 6(b) confirms this: LongVideo and Generation score 3.50 and 3.25, respectively, significantly higher than Storyboard's 2.25,Quality's 1.75, and Editing's 1.25. The results clearly show that within the current production chain, the capabilities of "continuous generation," "content evaluation," and "implementation" remain three distinct functions that have not yet been integrated into a stable closed-loop system. The dimension-stage coupling matrix and coupling contribution curves are shown in Figure 7.

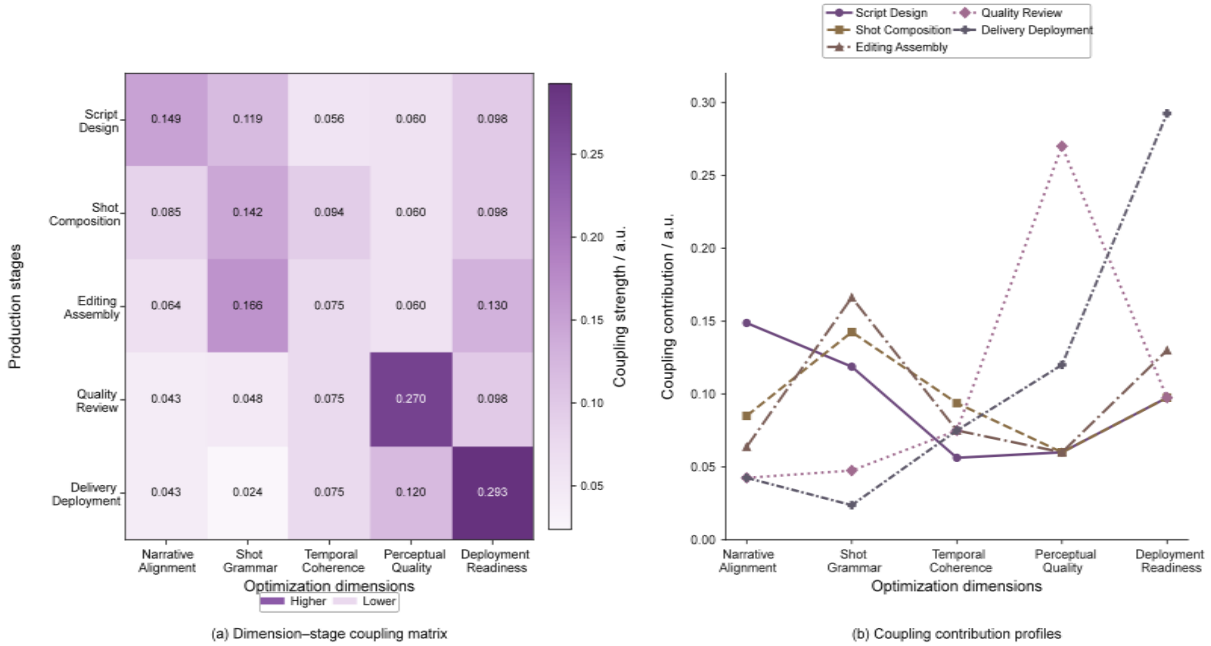

Figure 7: Dimension-Stage Coupling Matrix and Coupling Contribution Curve.

In Figure 7(a), the gap coefficients for the five dimensions are: Narrative Alignment 0.425, Shot Grammar 0.475, Temporal Coherence 0.375, Perceptual Quality 0.600, and Deployment Readiness 0.650. The largest gaps are not in temporal coherence, but in deployment readiness and perceptual quality.Figure 7(b) further presents the coupling contribution values for stage-level weaknesses. Among these, the coupling strength between Delivery Deployment and Deployment Readiness reaches 0.293, the highest among all combinations; the coupling strength between Quality Review and Perceptual Quality is 0.270, ranking second; the coupling strength between Editing Assembly and Shot Grammar is 0.166;Script Design and Narrative Alignment at 0.149; and Shot Composition and Shot Grammar at 0.143. This indicates that the relatively low scores in the current backend stages are not merely the result of a few categories dragging down the overall score, but rather directly correspond to two structural weaknesses in the underlying capability supply: one is insufficient deployment interfaces, and the other is insufficient quality feedback.

When these coupling relations are projected on actual failure modes, the problems hence become more intuitive. The most ordinary breakdown in the front end takes place when the semantic corresponding relation between the script and storyboards has already been built, but the produced output is short of high-quality feedback interfaces after entering the rough editing stage. This brings about that proposals are got up quickly, hence the final edition needs that much artificial hand polishing is done. The most frequent failing in the middle period is that although characters, actions, events still keep consistency, shot conversions and editing rhythms are not stable; the content can move forward, but the individual pictures themselves do not get enough establishment. The most representative malfunction in the rear end lies in that although quality estimations can recognize fuzziness, deformation, or text inconsistencies, they can not be directly transformed into proposals for amending, lens position adjustments, or editing route optimizations. Figures 6 and 7 together prove that the most urgent need currently is not to further lengthen the duration of a single generation operation or promote local continuity, but to arrange quality diagnostics, shot syntax, and deployment interfaces into a continuous sustainable feedback chain.

The verification of robustness also gives support to this conclusion. On the premise that the

total weight does not have any change, when we raise the weight of Deployment Readiness from 0.10 to 0.20, and lower the weights of Narrative Alignment and Perceptual Quality by 0.05 for each one, hence the following composite scores are got for the five technologies: Generation 1.20, LongVideo 1.10, Storyboard 1.05, Editing 0.91, Quality 0.80. The ranking has not undergone a fundamental reverse change; As has been pointed out, Generation still occupies the highest positions, LongVideo and Storyboard it is still they stay in the middle positions, and Editing and Quality are located in the bottom positions. The main transformation gets embodied in the compression of the middle positions: LongVideo is a little higher than Storyboard, therefore this shows that once engineering usability is returned to the center of the production chain, the value of continuous understanding and deployment interfaces becomes more prominent. This result makes clear that the conclusions which were talked about before in this paper do not rely on one alone weighting arrangement, but are decided by the real structure of the present technological ecosystem.

From the angle of deployment arrangement, the more realizable method on the current stage is not to seek complete-chain covering by one single model, but to arrange modules according to production stage: the front part uses Storyboard to define narrative limits and frame structure; the middle part utilizes Generation to make consecutive visual sections and Editing to revise the connections between shots; the long-process segment relies on LongVideo to maintain context and ensure continuous processing; and the back end uses Quality to perform quality diagnostics and feed back to the editing and generation interfaces.This phased configuration aligns with the data presented in Figures 5-7. The two areas truly worthy of priority enhancement at present are neither the initial script development nor the continuous generation in the middle phase, but rather the integration between quality review and delivery deployment. If these two stages continue to rely on manual intervention as a fallback, even the highest generation efficiency at the front end will struggle to consistently translate into final video quality and production efficiency.The above findings indicate that the next priority for AI-assisted film and television design and video production should shift from expanding individual capabilities to establishing a closed-loop backend workflow, integrating quality assessment, shot refinement, and deployment interfaces into a single workflow.

4 Conclusion

This paper addresses the issue of cross-stage content optimization in AI-assisted film and television design and video production by constructing a production-process-oriented analytical framework and comparing the stage-specific support capabilities of different technical approaches under a unified framework. Based on the results of evidence organization, structured coding, and stage mapping, the following three conclusions can be drawn.

(1) This present thesis has finished the arrangement of objects which are in accordance with the film and television making chain. Taking five technology categories—technology-storyboard pre-visualization, editing semantics, long-form video comprehension, controllable generation, and quality evaluation—as the core, we have built a staged evidence system that faces toward script design, shot composition, editing arrangement, quality checking, and delivery deployment. This makes former separate research things can be put into one comparison frame, thus it gives a united base for later stage-related assessments and picture-type analysis.

(2) The outcome shows that present technique abilities have clear features that belong to different stages. "Generation" and "LongVideo" have nearer positions to the core of cross-stage collaboration, "Storyboard" shows an obvious superiority in early-stage design, and "Quality" is mainly gathered in the back-end diagnosis stage. All in all, the support for front-end visual

display and middle-stage continuous content production is comparatively strong, hence quality feedback and distribution deployment still stay comparatively weak. This gives indication that the core bottleneck in movie and television design and video making has moved from "whether generation can be done" to "whether it can steadily get into the back-end closed loop."

(3) The limitation of this paper is that the analysis mainly bases upon evidence coding and structural comparisons; it has still not passed empirical testing verification under consistent video samples, a consistent video editing software environment, and a consistent subjective assessment protocol. The future research can further combine actual production projects to confirm the connection between stage scores and manual revision burden, final video quality, and delivery efficiency, and therefore promote the cooperative realization of quality diagnosis, shot optimization, and deployment interfaces in the identical working flow.

About the Author

Yi Wang was born in Zhangye City, Gansu, China in 1987. She graduated from Shanghai University with a master's degree. Currently, she works at Shanghai Pudong Vocational and Technical College, where her research focuses on film and television shooting and production. Aili Ren was born in Shanghai, China, in 1987. She received her master's degree from Shanghai University. She currently works at Shanghai Nanhu Vocational and Technical College. Her research interests include film and television shooting and production.

References

[1] Zhang, R., Yu, B., Min, J., et al. (2025). Generative AI for film creation: A survey of recent advances. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (pp. 6321-6333).

[2] Rao, A., Jiang, X., Guo, Y., et al. (2023). Dynamic storyboard generation in an engine-based virtual environment for video production. In SIGGRAPH 2023 Posters (pp. 40:1-40:2).

[3] Gu, X., Sun, Y., Ni, F., et al. (2023). TeViS: Translating text synopses to video storyboards. In Proceedings of the 31st ACM International Conference on Multimedia (pp. 4968-4979).

[4] Gu, X., Wang, X., Jin, C., et al. (2024). ScaMo: Towards text-to-video storyboard generation using scale and movement of shots. In Proceedings of the 6th ACM International Conference on Multimedia in Asia (pp. 115:1-115:8).

[5] Wei, Z., Wu, H., Zhang, L., et al. (2025). CineVision: An interactive pre-visualization storyboard system for director-cinematographer collaboration. In Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology (pp. 18:1-18:18).

[6] Argaw, D. M., Caba Heilbron, F., Lee, J.-Y., et al. (2022). The anatomy of video editing: A dataset and benchmark suite for AI-assisted video editing. In Computer Vision - ECCV 2022 (pp. 201-218).

[7] Stoll, E., Breide, S., Göring, S., et al. (2023). Automatic camera selection, shot size, and video editing in theater multi-camera recordings. IEEE Access, 11, 96673-96692.

[8] Li, B., Wu, Y., Lu, Y., et al. (2025). VEU-Bench: Towards comprehensive understanding of video editing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 13671-13680).

[9] Sandoval-Castañeda, M., Russell, B. C., Sivic, J., et al. (2025). EditDuet: A multi-agent system for video non-linear editing. In SIGGRAPH Conference Papers '25 (pp. 2:1-2:11).

[10] Fu, C., Dai, Y., Luo, Y., et al. (2025). Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 24108-24118).

[11] Wu, W., Liu, M., Zhu, Z., et al. (2025). MovieBench: A hierarchical movie-level dataset for long video generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 28984-28994).

[12] Ren, S., Yao, L., Li, S., et al. (2024). TimeChat: A time-sensitive multimodal large language model for long video understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 14313-14323).

[13] Song, Z., Wang, C., Sheng, J., et al. (2024). MovieLLM: Enhancing long video understanding with AI-generated movies. CoRR, abs/2403.01422.

[14] Yang, Z., Chen, D., Yu, X., et al. (2025). VCA: Video Curious Agent for Long Video Understanding. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 20168-20179).

[15] Zhang, H., Wang, Y., Tang, Y., et al. (2025). Flash-VStream: Efficient real-time understanding for long video streams. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 21059-21069).

[16] Jiang, Z., Han, Z., Mao, C., et al. (2025). VACE: All-in-one video creation and editing. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 17191-17202).

[17] Kara, O., Singh, K. K., Liu, F., et al. (2025). ShotAdapter: Text-to-multi-shot video generation with diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 28405-28415).

[18] Geng, D., Herrmann, C., Hur, J., et al. (2025). Motion prompting: Controlling video generation with motion trajectories. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1-12).

[19] Dalal, K., Koceja, D., Xu, J., et al. (2025). One-minute video generation with test-time training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 17702-17711).

[20] Lu, Y., Li, X., Li, B., et al. (2024). AIGC-VQA: A holistic perception metric for AIGC video quality assessment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (pp. 6384-6394).

[21] Qu, B., Liang, X., Sun, S., et al. (2024). Exploring AIGC video quality: A focus on visual

harmony, video-text consistency, and domain distribution gap. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (pp. 6652-6660).

[22] Zhang, Z., Jia, Z., Wu, H., et al. (2025). Q-Bench-Video: Benchmarking the video quality understanding of LMMs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3229-3239).

[23] Wang, J., Duan, H., Zhai, G., et al. (2025). AIGV-Assessor: Benchmarking and evaluating the perceptual quality of text-to-video generation with LMM. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 18869-18880).

[24] Madhusudana, P. C., Birkbeck, N., Wang, Y., et al. (2023). CONVIQT: Contrastive video quality estimator. IEEE Transactions on Image Processing, 32, 5138-5152.

[25] ITU-T. (2023). Recommendation ITU-T P.910 (10/2023): Subjective video quality assessment methods for multimedia applications. International Telecommunication Union.