Improving Video Generation with Human Feedback: A Comprehensive Survey

SurveyForge

Abstract— Human feedback has emerged as a critical component in advancing video generation systems, bridging the gap between computational efficiency and user-centric design. This survey provides a comprehensive overview of methodologies and frameworks that leverage human feedback to enhance video quality, relevance, and personalization. Key approaches include incorporating reinforcement learning from human feedback (RLHF) and integrating human insights into hybrid generative models, enabling iterative refinement and alignment with evolving preferences. The survey explores explicit and implicit feedback mechanisms, interactive video generation systems, and the role of community-driven inputs, highlighting their impact across domains such as entertainment, education, and marketing. It identifies challenges such as feedback variability, bias, and scalability while emphasizing the role of sophisticated evaluation frameworks like VBench in aligning automated metrics with human perceptions. Future directions include refining multi-modal feedback systems, developing real-time adaptation mechanisms, and fostering inclusivity in human-aligned generative systems. By synthesizing user input with advanced machine learning methodologies, the integration of human feedback in video generation offers transformative potential for creating rich, engaging, and highly personalized content.

Index Terms—human feedback integration, video generation frameworks, personalized content creation

1 Introduction

The significance of enhancing video generation through human feedback lies at the confluence of advanced technological developments and the evolving paradigms of user engagement. As video generation technologies have transitioned from traditional methodologies anchored in manual editing to modern, AI-based frameworks, the necessity for a more interactive approach to content creation has become evident. Human feedback serves as a vital component in this context, infusing generative frameworks with user-centric insights that help refine and enhance outputs in terms of quality, relevance, and engagement.

Historically, video generation has struggled with quality assessment due to the inherent complexities of evaluating dynamic visual media. Traditional assessment techniques primarily focused on visual fidelity without adequately addressing temporal coherence and user satisfaction. Recent advancements, however, demonstrate shifts towards integrating generative adversarial networks (GANs) and deep learning approaches that leverage the strengths of human evaluators. For instance, GANs, as discussed in [1], synthesize video content that is closer to real human expectations by learning from pairs of input-output examples curated from user feedback, thus providing a substantial avenue for enhancing qualitative output.

The use of human feedback mechanisms is multifaceted, encompassing explicit and implicit dimensions. Explicitly, users can provide direct assessments and modifications, allowing for immediate adjustments in the video generation pipeline. Implicit feedback mechanisms such as engagement metrics (e.g., viewing time, replays) can autonomously optimize video content based on real user interactions. The integration of these feedback channels has shown to lead to significant improvements in content personalization, as demonstrated in studies that utilized crowd-sourced platforms for iterative model training [2].

Yet, this integration is not without challenges. One of the key obstacles in effectively harnessing human feedback is the inherent variability in user preferences. Factors such as cultural context, individual taste, and differing interpretation of visual cues can introduce inconsistency in feedback. Research highlights that a systematic, fine-grained collection of human assessments via structured methodologies can greatly enhance the learning process for AI systems [3]. For example, the adoption of reinforcement learning frameworks can effectively model human preferences in a structured format, thus allowing for the continuous adaptation of generative models to align with evolving user standards.

Emerging trends within the field underscore the importance of community and user interactions in shaping video generation technologies. Collaborative platforms and participatory design methods provide avenues for continuous improvement, allowing for real-time feedback and fostering a sense of shared ownership over the creative process. These approaches challenge the conventional paradigm of content generation, position users as co-creators, and create an iterative cycle of feedback and enhancement that is crucial for long-term engagement.

Furthermore, the complexity of human cognition presents an exciting challenge and opportunity for future advancements. As researchers begin to explore the realms of explainable AI and augmented reality, the potential for deeply personalized video generation methods utilizing human feedback becomes increasingly tangible. Such advancements can potentially revolutionize applications across varying domains, from personalized marketing to instructional content. Thus, optimizing the integration of human feedback into video generation frameworks not only enhances the quality and relevance of generated videos but

also aligns the technology with broader societal needs.

In conclusion, the integration of human feedback into video generation represents a critical advancement that bridges the gap between human creativity and machine capabilities. The exploration of adaptive frameworks that can incorporate nuanced human insights will represent the frontier of research and application in this rapidly evolving domain. As the field progresses, the ability to generate content that resonates with audiences will be paramount, ultimately driving innovation and engagement across various sectors while emphasizing the necessity for ethical considerations in AI-generated content.

2 Frameworks for Integrating Human Feedback

2.1 Human-Centric Reinforcement Learning Frameworks

The integration of human feedback within reinforcement learning (RL) frameworks has emerged as a transformative approach in enhancing video generation systems. This section explores several key methodologies that harness human evaluative signals to refine the training of generative models, thereby improving output quality and user satisfaction. Central to these frameworks is the premise that human preferences can offer richer, more nuanced guidance than traditional predefined reward structures.

One prominent example of human-centric reinforcement learning is the TAMER framework, which encompasses a paradigm shift by allowing agents to learn from human feedback instead of relying solely on manually designed rewards. In TAMER, an agent receives direct feedback from humans on its actions, updating its behavior dynamically to maximize alignment with human expectations. This approach has been shown to outperform traditional RL methods in tasks requiring adaptability to complex human preferences, particularly in environments where feedback richness is crucial, such as video generation [4].

Building on the principles of TAMER, the COACH framework introduces a structured learning approach that leverages human critiques and evaluative signals during training. COACH employs an architecture that integrates a user feedback model to predict the effectiveness of generated video segments, thereby enabling the reinforcement learning agent to adaptively refine its policies based on this user input. This model emphasizes the importance of iteratively incorporating feedback, thereby enhancing the agent's ability to produce coherent and contextually relevant videos [5].

Another innovative approach is FRESH (Feedback-based Reward Shaping), which utilizes continuous human feedback during training to fine-tune agent policies. By shaping the reward function based on real-time user interactions, FRESH demonstrates how iterative engagement can lead to more streamlined optimization paths for generative tasks, significantly improving the quality of video outputs. This strategy not only accelerates learning but also ensures that the generated content is closely aligned with user expectations [2].

The potential of deep RL methods, particularly when combined with human feedback, further highlights the advantages of integrating user interactions. By incorporating corrective signals from human evaluators, agents can effectively navigate complex decision landscapes and fine-tune behaviors that align with nuanced human preferences. This methodology has been illustrated through various applications, where human feedback effectively guides exploration in high-dimensional spaces, facilitating the generation of engaging video narratives that resonate with viewers [6].

While the benefits of these frameworks are substantial, there are inherent trade-offs and challenges to consider. The reliance on human feedback introduces variability that may complicate training dynamics, such as the risk of overfitting to specific user preferences at the expense of broader applicability. Moreover, the quality and reliability of the feedback mechanisms themselves pose obstacles; systems must effectively filter out biased or inconsistent feedback to ensure robust model training. Emerging methods aim to address these issues through structured feedback mechanisms that enhance the reliability of human evaluations, including multimodal feedback collections that capture diverse user input [7].

Looking toward future directions, the continuous evolution of human-centric RL approaches offers promising avenues for enhancing video generation. The incorporation of fine-grained feedback systems, where users can provide detailed critiques on specific aspects of generated videos, could lead to even more refined model adjustments. Additionally, the development of platforms for collecting extensive datasets of user interactions will further bolster the training of RL models, enabling them to generalize better across varied content generation tasks. As the landscape evolves, leveraging human feedback in reinforcement learning is likely to play a pivotal role in shaping the next generation of intelligent video generation systems.

2.2 Hybrid Generative Models with Human Feedback

Hybrid generative models that integrate traditional techniques with human feedback show considerable promise in enhancing video generation processes, particularly in terms of creativity and content alignment. By leveraging the strengths of established frameworks such as Generative Adversarial Networks (GANs), these models incorporate user-derived insights to inform and refine generative processes. This integration addresses key limitations inherent in traditional generative models, such as the challenges of aligning output with subjective human intentions and the constraints posed by static training datasets.

At the core of these hybrid models is the architecture of GANs, comprising a generator and a discriminator that engage in a continuous interplay to create and evaluate realistic outputs. Incorporating human feedback into this framework can significantly enhance the generator's ability to produce content that resonates with users, as human evaluative signals serve as a guiding force for the generator's outputs. For instance, Creative Adversarial Networks (CAN) utilize human ratings to improve not only the fidelity of generated videos but also to encourage creative deviations from conventional styles, thus fostering innovation in content generation [6].

Various methodologies exemplify the effective integration of human feedback within GAN structures. A particularly promising approach involves augmenting the discriminator with a human feedback mechanism that provides real-time evaluative signals during training. This strategy enables the discriminator to assess not just the realism but also the subjective quality of the generated content. Systems employing this method have demonstrated improved alignment with human preferences, yielding outputs that are both statistically and aesthetically superior to those produced by conventional models [8]. Furthermore, this integration has the potential to alleviate the common challenge of mode collapse within GANs, where the generator produces limited variability in outputs.

Design considerations for hybrid models necessitate the establishment of effective feedback structures. Human feedback can be multidimensional—ranging from simple binary classifications of 'good' versus 'bad' to more nuanced evaluations encompassing attributes such as aesthetic appeal, coherence, and engagement [9]. Formally, we can express the feedback signal as a function $F(v)$, where v denotes a video frame created by the model. The objective becomes maximizing the generation process to enhance expected feedback:

$$\max_{\theta} E\,[10] \quad \text{where } v = G(z;\theta)$$

In this formulation, $G(z;\theta)$ represents the generative process conditional on the latent variable z and parameters θ.

Future research should also address the trade-offs between relying on objective metrics versus subjective human evaluations, as the latter may introduce variability due to individual biases and personal preferences. Sole reliance on human feedback could complicate the model's generalization capabilities since feedback dynamics may differ across various viewer baselines and contexts. Understanding the implications of the contexts in which human feedback is elicited—such as user expertise and cultural background—is crucial for accurate interpretation and integration into generative frameworks [11].

Emerging trends in using reinforcement learning-based techniques for hybrid generative models open new avenues for exploration. These methods focus on fine-tuning by continuously adapting to user feedback gathered over multiple interactions, fostering a dynamic feedback loop that refines both input processes and generated outputs [12].

In conclusion, as hybrid generative models continue to evolve, critically examining the integration of human feedback will lead to breakthroughs in achieving greater creativity and relevance in video generation tasks. Future research should aim not only to optimize feedback mechanisms but also to develop robust evaluation protocols that balance human input with automated assessments, ensuring models are both effective and adaptable to diverse user landscapes.

2.3 Interactive User-Guided Video Generation

Interactive user-guided video generation frameworks represent a significant evolution in content creation, allowing users to provide real-time feedback and make adjustments throughout the video generation process. This paradigm empowers users by enhancing their involvement in the creative workflow, thus facilitating a more personalized and dynamic video production experience. By leveraging user inputs through various means—such as textual commands, visual prompts, or even drawing specifications—these frameworks enable a collaborative synergy between human creativity and machine intelligence.

Several approaches to interactive video generation have emerged, each with unique strengths and limitations. For instance, the InteractiveVideo framework integrates multimodal instructions, permitting users to control video generation through diverse input methods, from text prompts to visual modifications, facilitating a highly personalized creation process [13]. This flexibility enhances user engagement by enabling iterative refinements, allowing for real-time edits that substantially impact the final video output.

Another notable example is the VideoMotion customization framework, which utilizes temporal attention mechanisms to adapt the generated content based on user-defined motion parameters. This method enables users to specify desired movements while ensuring that the generated videos remain coherent and visually appealing. The synergy between user specifications and machine learning enhances the realism and specificity of the generated content, but also introduces complexity in accurately deciphering user intents without oversimplifying the motion dynamics [14].

However, these frameworks can present trade-offs in their design and implementation. A key challenge faced is the balance between user control and the underlying generative model's stability. For instance, excessive user influence may lead to inconsistent video outputs if the model cannot generalize adequately across a vaunted array of user specifications. Further, real-time feedback loops require efficient processing capabilities—limiting the complexity of models that can be realistically deployed in interactive environments [15].

Moreover, these interactive frameworks often rely on user annotations that can be subjective, thereby impacting the model's learning and performance. As noted in the study on user feedback integration, incorporating diverse user inputs can enrich generated content, but also highlights the difficulty of reconciling differing aesthetic preferences across users [16]. Thus, while striving for a user-guided model, there remains a necessity to implement robust mechanisms for normalizing and interpreting the feedback to mitigate individual biases.

Emerging trends in this space include harnessing generative adversarial networks (GANs) for more sophisticated interactive capabilities, where users can manipulate features iteratively in a semantic latent space. The GANs' inherent structures lend themselves toward flexibility in generative control, yet their application in real-time feedback loops raises questions regarding the scalability of training processes and the quality of generated results across various datasets [1]. Furthermore, integrating recent advances in reinforcement learning can facilitate training dynamic models that adapt not only to explicit user feedback but also implicitly inferred preferences through patterns in user interaction [17].

In terms of future directions, further research is needed to enhance the interpretability of user commands within generative frameworks. This could include improving model designs to better accommodate multi-dimensional feedback, fostering an environment where real-time adjustments can be made seamlessly and effectively. The integration of advanced human-in-the-loop mechanisms and hybrid approaches combining explicit user input alongside implicit feedback mechanisms will likely enhance the robustness and versatility of interactive video generation systems. This evolution will necessitate the development of standardized evaluation metrics to accurately assess user satisfaction and the overall quality of generated videos, ensuring that these innovations align well with user expectations and creative intent.

2.4 Frameworks for Implicit Feedback Mechanisms

Models leveraging implicit human feedback through observed user behavior and engagement metrics represent an evolving approach in the domain of video generation. The capacity to infer user preferences from indirect signals allows for dynamic adaptation of generation models, ultimately enhancing overall performance while reducing the reliance on explicit user input. These systems typically analyze user interactions—such as viewing time, engagement rates, and pattern recognition—to fine-tune generation processes effectively.

Engagement analytics tools, which assess viewer interactions during video playback, facilitate the extraction of preference signals without requiring users to provide direct feedback. Metrics such as average watch time, repeat views, and user shares serve as effective indicators of content preference. For instance, the work by Zhang et al. [18] illustrates how implicit feedback mechanisms guide the refinement of generative models based on user interaction metrics. In this framework, these metrics are pivotal in shaping content production, aligning it with audience expectations.

One of the core strengths of implicit feedback mechanisms lies in their scalability. By gathering data passively without additional user burden, they continuously inform models, even in heterogeneous user environments. This offers a significant advantage over explicit feedback systems, which may suffer from biases due to limited user participation. However, a challenge persists: the accuracy of inferred user preferences can be skewed by confounding variables, such as external distractions or the context of viewing, as evidenced by varying engagement levels across different demographic groups [19].

Moreover, implicit feedback systems play a crucial role in the creation of personalized content. By analyzing user behavior patterns, algorithms can adjust generative parameters to cater to individual preferences effectively. For instance, models can prioritize certain content themes that statistically resonate more with specific audience segments. The work by Chen et al. [20] exemplifies this principle, demonstrating how video generation models adapt outputs based on aggregated viewer interaction metrics, thereby significantly enhancing viewer satisfaction and engagement.

Mathematically, models often employ reinforcement learning strategies to optimize the impacts of implicit feedback. This involves defining a reward function that translates user engagement metrics into numerical assessments that guide model adjustments. Let r represent the reward signal derived from implicit user feedback, which could be computed as a function of various engagement metrics. The aggregation of these signals can be expressed as:

$$R = \sum_{t=1}^{T} f(E_t),$$

where $f(E_t)$ denotes the function mapping engagement metrics E_t observed at time t, and T corresponds to the observation duration. Through this approach, models undergo iterative updates that help align them more closely with user interaction data.

Despite their enormous potential, implicit feedback mechanisms are not without limitations. A significant concern is the potential for misinterpretation of engagement metrics. High engagement does not always equate to a positive viewer sentiment; thus, systems must be adept at distinguishing genuine interest from mere curiosity or habitual viewing [18]. Additionally, there's a risk of overfitting models to specific user behaviors observed during training phases, which could constrain their adaptability to new audience segments.

Emerging trends point towards a shift toward hybrid systems that combine explicit user feedback with implicit metrics to enhance accuracy. By integrating both feedback types, models can achieve a more holistic understanding of user preferences. This synthesis is particularly emphasized in recent studies aimed at learning from dynamic user interactions across platforms, paving the way for more robust and nuanced video generation systems [18], [21].

In conclusion, while frameworks for implicit feedback mechanisms offer innovative pathways for enriching video generation processes, addressing their inherent challenges remains critical. Future research should focus on refining methodologies for accurately inferring user preferences and developing adaptive systems capable of responding meaningfully to inferred feedback. This approach will ultimately enhance audience engagement and satisfaction in video content creation, complementing the ongoing exploration of human feedback integration in generative frameworks.

2.5 Methods for Evaluating Human Feedback Integration

Evaluating the integration of human feedback into video generation models is vital for optimizing performance and aligning outputs with user expectations. Various approaches exist, each with unique methodologies and implications for assessing the effectiveness of feedback incorporation. These can be broadly categorized into qualitative and quantitative evaluation metrics, underscoring the need for a nuanced understanding of how human feedback influences generative outcomes.

A widely adopted method is the use of subjective quality metrics, such as the Mean Opinion Score (MOS), where human evaluators rate generated videos based on predefined criteria like coherence, realism, and engagement. This approach has been exemplified in studies where traditional

scoring has been compared with user preferences, revealing discrepancies that automated metrics may not capture effectively. Notably, the observed correlation between human ratings and automatic evaluation scores often fluctuates, underscoring the potential limitations of solely relying on objective measures [22].

On the quantitative side, objective metrics such as Fréchet Video Distance (FVD) and Video Quality Index (VQI) focus on measurable attributes of generated videos—like temporal stability and visual fidelity—providing a rigorous framework for assessing performance against diverse datasets. Unlike subjective metrics, which can be influenced by personal biases, these objective measures allow for scalable comparisons across different models and conditions. Nonetheless, the adoption of such metrics also presents challenges; for instance, models trained under constrained circumstances might yield high scores even while producing poor content, as indicated by empirical studies on performance over optimization using mislabeled or biased datasets [6].

Emerging trends in evaluation methodologies highlight the necessity of combining qualitative and quantitative evaluations to produce a holistic view of model performance. One such innovative approach involves fine-grained feedback mechanisms, allowing human evaluators to provide detailed critiques on specific attributes of generated outputs [3]. This method enhances the richness of the feedback, enabling the system to correct flaws effectively—yet it demands significant interaction from users, which could be impractical at scale.

Moreover, the use of reinforcement learning algorithms tailored to funnel human feedback into model adjustments, such as the Human Preference Score framework, demonstrates a shift toward more dynamic evaluation processes. By calibrating feedback to attack specific aspects of generation, models become more adept at addressing user dissatisfaction directly and improving iteratively [23]. Interestingly, these approaches emphasize the critical balance between operator effort and automated evaluation efficiency, suggesting a roadmap for future research focused on optimizing interaction costs while maximizing output quality.

Challenges remain in the integration of human feedback into existing evaluation frameworks, particularly regarding the scalability and reliability of human evaluations in large datasets. While current solutions often rely on qualitative assessments, incorporating adaptive algorithms capable of interpreting and responding to human input in real-time could streamline feedback integration further [24]. As such, developing robust datasets that mirror real-world audience interaction remains a priority, with recent trends suggesting potential in harnessing software-generated data as reliable proxies for human feedback.

In summary, the multifaceted nature of evaluating human feedback integration in video generation models encompasses diverse methodologies that span subjective ratings, objective assessments, and innovative feedback mechanisms. Future avenues should explore the synthesis of these methods while addressing inherent challenges of scalability, biases, and variability in human judgment—ultimately driving improvements towards user-centric video generation frameworks.

3 Human Feedback Mechanisms and Data Collection

3.1 Direct Feedback Mechanisms

Direct feedback mechanisms are integral to the process of enhancing video generation systems, as they allow users to provide explicit evaluations and suggestions regarding generated content. These mechanisms, encompassing a range of approaches such as surveys, rating systems, and comment sections, facilitate the collection of user preferences and critiques, thereby providing insights that influence the refinement of generative models.

Surveys and questionnaires are commonly employed to gauge viewer satisfaction and identify areas needing improvement. They can be meticulously designed to include specific queries about various aspects of video, such as narrative coherence, aesthetic value, and emotional impact. For instance, a well-structured survey may involve Likert scale responses, enabling granular evaluations of video elements. Insights gleaned from such tools can significantly enhance model training, as user inputs directly inform generative processes, yielding outputs that better align with viewer expectations. Research has shown that user-expressed feelings through surveys can yield qualitatively stronger generative models than those trained on numerical data alone, highlighting the effectiveness of direct human feedback in refinement processes [2].

Rating systems provide an alternative, offering a more streamlined approach to feedback collection. Simple rating scales, such as star ratings from 1 to 5, allow users to quickly convey their impressions of video quality and design. The primary advantage of this approach lies in the ease of data processing, as numerical ratings can be aggregated and analyzed statistically to discern patterns in user preferences. However, a significant limitation is that ratings often lack the nuance found in qualitative feedback, leading to potential oversimplification of critical insights [1]. Furthermore, reliance on scales can bias user responses, as individuals may tend to avoid extreme ratings, which complicates the analysis of truly negative or positive experiences.

Comment sections implemented alongside video outputs serve as an invaluable tool for unfiltered user expression. These open channels facilitate detailed, qualitative feedback that can reveal deep-seated viewer sentiments about specific video elements. Users can convey their likes and dislikes, facilitating more nuanced revisions of generative strategies. Such qualitative insights can be particularly useful for creators seeking to understand subjective viewing experiences, though this feedback can also be overwhelming due to the unstructured nature of comments [25].

Emerging trends in direct feedback mechanisms spotlight the necessity for incorporating more dynamic and interactive strategies. Techniques that leverage real-time feedback interactions allow users to make adjustments during video consumption. For example, systems that enable viewers to modify parameters such as pacing or scene selection while the video plays elevate user engagement and satisfaction. Such interactive mechanisms challenge traditional static models of user feedback and suggest a move towards adaptive generative systems capable of tailoring outputs to immediate user preferences [3].

Despite the advantages of direct feedback mechanisms, several challenges persist. The reliability of user feedback is often subject to variations in individual tastes, biases, and external factors influencing user experiences. Moreover, there remains an ongoing risk of noise in feedback data, necessitating sophisticated filtering mechanisms to separate valuable insights from irrelevant noise. This challenge is amplified by the need for models to adapt to a diverse user base, which requires significant computational resources and rigorous quality assessment processes.

In conclusion, direct feedback mechanisms play a pivotal role in refining video generation systems. As systems evolve, future research should emphasize the development of hybrid strategies that integrate qualitative and quantitative feedback, leveraging advances in natural language processing to analyze user comments effectively. Furthermore, building robust infrastructures that support real-time, interactive feedback will likely open new avenues for enhancing user experience and model performance, fortifying the symbiotic relationship between user preferences and generative output quality [5].

3.2 Implicit Feedback Mechanisms

Implicit feedback mechanisms leverage user interactions and engagement metrics to infer preferences and enhance video generation quality without requiring explicit input from users. These mechanisms focus primarily on observing viewer behavior, such as watch time, click-through rates, replays, and dropout points, which can yield valuable insights into what aspects of the video resonate most with the audience. This data-driven approach presents a scalable and efficient alternative to traditional explicit feedback methods.

One prominent technique utilized in implicit feedback mechanisms is engagement tracking. By quantifying metrics like total viewing time and replay rates, systems can identify which video segments capture interest or lead to disengagement. For example, videos with longer average watch times and higher replay rates indicate content that effectively engages viewers. This feedback loop allows for iterative content refinement based on direct viewer behavior, enabling systems to automatically adjust elements of video generation to align with user preferences. The ability to analyze viewer interactions in real time not only enhances current outputs but can also contribute to more informed long-term strategies for video content creation, as demonstrated in the assessment of various audience engagement patterns across different formats [26].

Complementing engagement tracking is behavioral analysis, which examines specific user interactions with video content. Metrics like heatmap analysis can visualize viewer engagement across different segments of a video. For instance, sections with a high frequency of rewinding or replaying may be recognized as particularly impactful, guiding content creators to focus on similar elements in future videos [27]. Additionally, dropout rates—points at which viewers discontinue watching—highlight areas that may require further refinement due to perceived irrelevance or lack of engagement. By systematically analyzing these behavioral patterns, video generation models can learn to produce content that aligns more closely with user expectations.

Despite their numerous advantages, implicit feedback mechanisms do present limitations. One major challenge is the interpretation of data derived from purely observational metrics. Without explicit input, there's a risk of misattributing user behavior based on flawed assumptions about viewer preferences. For example, a high drop-off rate may not solely indicate unappealing content but could also stem from external factors such as competing viewer commitments or technical issues [28]. Furthermore, engagement metrics may inadvertently favor certain genres or styles, potentially leading to an echo chamber effect where diversified content is underrepresented.

Emerging trends in integrating artificial intelligence with implicit feedback mechanisms are noteworthy. Recent advancements in reinforcement learning from human feedback (RLHF) are particularly relevant, offering a pathway for systems to evolve based on implicit signals. Techniques such as the "Inter-temporal Bradley-Terry" model can effectively capture human judgments, even in systems trained primarily through implicit data, resulting in a more nuanced understanding of user interaction patterns [29]. Additionally, hybrid methods that synthesize both implicit and explicit feedback can address some of the limitations associated with relying solely on engagement metrics.

Looking to the future, innovative potential in this domain lies in developing robust algorithms capable of transforming complex user behaviors into actionable insights while mitigating biases inherent in user engagement data. This involves exploring multi-modal analytics, integrating visual and textual user feedback, and dynamically adjusting content in response to shifting viewer preferences over time. As implicit feedback mechanisms continue to evolve, they stand to play a pivotal role in enabling adaptive, user-centered video generation systems that resonate more deeply with audiences. By effectively harnessing these implicit signals and feeding them into adaptive learning processes, the future of video generation could be characterized by an unprecedented level of personalization and alignment with viewer expectations, marking a significant advancement in the relationship between content generation and user experience.

3.3 Community-driven Feedback Systems

Community-driven feedback systems leverage collective insights from user communities to enhance video generation models through iterative refinements, making the generated content more aligned with user preferences. These systems capitalize on the multiplicity of perspectives found within a diverse user base, allowing for richer feedback than any single user could provide. This subsection explores the mechanisms and frameworks underlying such systems, highlighting their strengths, limitations, and the potential for future developments.

One primary approach in community-driven feedback systems involves crowdsourced evaluation platforms where users are invited to review and evaluate generated video outputs. Such platforms enable a broad range of evaluative opinions, fostering diversity in feedback that mitigates biases present in individual critiques. For example, methods that employ structured feedback from users, much

like crowdsourcing techniques in content creation, leverage large participant pools to assess video quality broadly. Platforms like these have been shown to effectively reduce individual biases and create a more nuanced understanding of user preferences, ultimately leading to refined outputs that cater to a wider audience [30].

Further, user forums and discussion boards amplify community engagement, allowing users to share their experiences and insights regarding generated content. Facilitating rich discussions around the strengths and weaknesses of videos enables a collaborative environment where users can brainstorm and propose improvements. This setup supports iterative refinement of video generation models, as ideas generated in these discussions can directly influence subsequent model training cycles. For instance, the aggregation of qualitative feedback gathered through forums can inform model adjustments that might not have been apparent through isolated user feedback alone [31].

Moreover, rating aggregation mechanisms serve as a method to combine diverse user evaluations into a cohesive assessment of video quality. By employing statistical techniques to weight ratings based on user reliability and previous evaluations, platforms can derive representative scores that reflect community sentiment accurately. This approach helps in determining the overall quality and user satisfaction with generated videos, amplifying the effects of individual critiques by treating them as component parts of a larger evaluative framework. Implementing methodologies such as mean opinion score aggregation underlies much of this interpretation, providing clear insights into how collective feedback correlates with refined generation outcomes [15].

However, community-driven feedback systems face certain limitations. The sheer volume of feedback can complicate the integration process into model training workflows. Many systems struggle with efficiently parsing and prioritizing feedback, especially when faced with conflicting opinions. This may necessitate the development of sophisticated algorithms capable of discerning actionable insights from the noise of large data sets while maintaining the responsiveness that characterizes successful user engagement [7]. Furthermore, as these systems are not free from biases, ensuring equitable representation in feedback collection remains a challenge. Factors such as demographic skewing in user populations can lead to generative models that are poorly aligned with broader societal norms and preferences [32].

Emerging trends also indicate a push towards implementing machine learning techniques to enhance the analysis of community feedback. For example, developing predictive models to forecast user preferences based on historical feedback could greatly enhance the personalization of video generation systems. Such advancements suggest the integration of advanced natural language processing techniques to interpret qualitative feedback dynamically.

In summary, community-driven feedback systems present a promising avenue for refining video generation methodologies by utilizing the collective wisdom of diverse user bases. As these systems evolve, developing mechanisms to effectively integrate feedback while addressing inherent biases and optimizing for relevance and scalability will be crucial for maximizing their effectiveness in enhancing user satisfaction and engagement in synthesized media. Future directions should focus on refining algorithms for feedback processing and exploring the integration of AI-driven models that can synthesize varying perspectives into coherent adjustments in generative outputs, thereby truly embodying the collaborative potential of community-driven systems.

3.4 Feedback Integration into Model Training

Integrating human feedback into the training processes of video generation models represents a transformative approach aimed at aligning model outputs more closely with user expectations. This integration encompasses several methodologies that focus on capturing and leveraging real-time user evaluations to refine the generative capabilities of models. The prevailing paradigms within this domain can be broadly categorized into reinforcement learning frameworks that utilize human evaluative signals, direct feedback mechanisms that inform model adjustments, and hybrid approaches that synthesize both automated and human-critical evaluations.

A prominent methodology in this space is the Reinforcement Learning from Human Feedback (RLHF) paradigm, which adapts Q-learning frameworks to incorporate human feedback as a form of reward signal. This training strategy enables models to dynamically adjust based on user preferences. For instance, RLHF channels human evaluations into a reward mechanism that explicitly optimizes model parameters, enhancing the realism and relevance of generated videos. Evidence from works like "Training language models to follow instructions with human feedback" demonstrates improved model outputs when human feedback is integrated into training via rewards derived from user ratings of output quality [33].

In contrast, direct approaches leverage explicit feedback from users, such as surveys and ratings, during training iterations. This method can yield immediate insights into how well generated content meets user expectations. Rating systems, for example, facilitate quick assessments of various aspects of video quality. Studies, such as those described in "Responsive Action-based Video Synthesis," have shown the success of this method, where user ratings directly influenced the iterative refinement of generative models, aligning outputs more closely with user desires [34]. However, reliance on a limited set of user inputs in this approach can introduce biases, highlighting the need for a broader data collection strategy to ensure diversity in feedback.

Another innovative strategy involves the incorporation of crowd-sourced evaluations into a feedback loop, allowing models to learn from a diverse array of human preferences. This model has been effectively leveraged in applications like scenario-based content generation, where extensive feedback datasets from various users provide rich insights for model training. The advantages of this approach are clearly articulated in the findings from the paper "Using the Crowd to Generate Content for Scenario-Based Serious-Games," illustrating how the diversity of input can craft a more adaptable content generation framework [35]. However, the variability in output quality associated with crowd-

sourced feedback necessitates the implementation of robust filtering mechanisms to enhance data quality.

Emerging trends in feedback integration highlight a shift toward combining different methodologies to harness the strengths of each. For instance, frameworks that incorporate both implicit feedback—derived from user engagement metrics—and explicit feedback are becoming increasingly valuable. Strategies that dynamically adjust the weights of feedback based on its source and reliability can enhance the adaptability of training algorithms while addressing concerns over biased data inputs.

Moreover, emerging technologies such as generative adversarial networks (GANs) facilitate more sophisticated integrations of feedback into the training process. Approaches like that described in "Deep Interactive Evolution" propose interactive evolution methods where human users guide the generator's latent space, effectively co-creating content that aligns with their artistic intents [36]. This innovative interplay enables a more creative and responsive model capable of producing high-quality outputs that reflect complex human preferences.

Despite these advancements, challenges persist, particularly regarding the scalability of human feedback mechanisms. As the volume of generated data increases, maintaining an efficient feedback loop that processes user evaluations while preserving output quality becomes increasingly complex. Additionally, the ethical implications surrounding privacy and consent in data collection must be navigated carefully to ensure responsible integration of user feedback into training regimes.

In summary, the integration of human feedback into model training paradigms for video generation presents both significant opportunities and challenges. As methodologies continue to advance, a balanced approach that appropriately weighs different types of feedback while employing robust mechanisms to filter and refine input will be crucial for developing models that are not only technically adept but also aligned with user expectations and societal norms. The evolution of hybrid feedback systems is poised to characterize the future landscape of video generation, establishing them as vital components of user-centered design in AI.

3.5 Evaluation Metrics for Feedback Analysis

Human feedback plays a critical role in the iterative refinement of video generation systems, and a robust evaluation framework is essential to assess its effectiveness. Evaluating the impact of human feedback on generated videos necessitates a dual approach combining subjective measures, which capture human perceptions of quality, and objective metrics, which provide quantifiable performance indicators. Subjective metrics, such as the Mean Opinion Score (MOS) and other user-centric evaluations, are predicated on directly soliciting ratings from human observers regarding various quality aspects—such as coherence, engagement, and relevance. For example, in multiple studies, MOS has proven effective in establishing a baseline for perceived video quality by leveraging human ratings across diverse demographics and contextual settings [33].

On the flip side, objective metrics such as Fréchet Video Distance (FVD) and Video Quality Index (VQI) are designed to provide consistent, quantitatively rigorous assessments of video outputs. FVD, in particular, measures the distance between the distributions of real and generated video features using a pre-trained deep learning model, thus capturing both spatial and temporal dynamics which are challenging to evaluate solely through subjective assessments [37]. The merits of objective metrics lie in their ability to provide quick, repeatable evaluation results that reflect coherent video generation patterns across extensive datasets, contrasting with the often time-intensive nature of human evaluations.

Nevertheless, the relationship between subjective and objective quality measures can be tenuous. Several studies have identified discrepancies where high objective scores do not necessarily correlate with favorable human perceptions. This highlights the necessity for a comprehensive evaluation framework that harmonizes both subjective and objective measures to foster more holistic assessments of video generation quality. For instance, combining user feedback with objective metric outcomes—in a multi-criteria rating system—may enable deeper insights into the nuanced preferences of users as evidenced through qualitative user feedback surveys [38].

Emerging methodologies such as engagement analytics offer pathways to enhance feedback evaluation frameworks. By leveraging implicit feedback mechanisms through observed user interactions—like watch time and click-through rates—models can glean insights into viewer preferences without requiring explicit ratings [39]. Moreover, crowdsourced evaluations, where collective user feedback is synthesized, can significantly sharpen the reliability of generated assessments. Though such community-driven frameworks may introduce variability in feedback consistency, they capture a broader range of preferences, which is crucial for developing algorithms that generalize well across variably engaged user groups [24].

However, this integration brings challenges, particularly regarding the potential biases in crowd-sourced feedback and the inherent quality of user data. It is paramount to establish robust protocols that address these limitations, ensuring the evaluative processes remain equitable and reflective of diverse user experiences. Furthermore, innovative approaches like reinforcement learning from human feedback (RLHF) and adaptive sampling methods, which utilize user-generated feedback to optimize model parameters dynamically, point toward fruitful research avenues. These methods—by integrating real-time human insights into model training iterations—could enhance the practical applicability of video generation systems in real-world contexts [40].

In synthesis, developing an evaluative framework that accommodates both subjective judgments and objective metric assessments represents a pivotal step towards refining the quality of video generation systems through human feedback. This dual-focused evaluation paradigm is likely to foster advancements in video generation technologies, enabling more responsive and intuitive systems. Moreover, as feedback methodologies continue to evolve, ongoing research should focus on standardizing evaluation metrics to better align automated assessments with human perceptions, ensuring that future advancements resonate closely

with user expectations and needs. Such considerations could transform the landscape of human feedback integration in video generation, enhancing user engagement while maintaining technological rigor.

4 Evaluation Metrics for Human Feedback Integration

4.1 Understanding Video Quality Metrics

Video quality metrics are essential in evaluating generated video outputs, serving as benchmarks for assessing both realism and coherence. A comprehensive understanding of these metrics not only informs the effectiveness of generative models but also aligns their outputs with human expectations. Various metrics can be categorized into objective measures, which are algorithmically derived, and subjective measures, which rely on human judges.

Traditional video quality metrics such as Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) focus primarily on visual fidelity by quantifying the perceptual differences between the generated video and a reference video. PSNR calculates a logarithmic scale of the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation. Given by:

$$PSNR = 10 \cdot \log_{10} \left(\frac{R^2}{MSE} \right)$$

where R is the maximum possible pixel value and MSE is the mean squared error, PSNR remains widely applied due to its simplicity. However, its sensitivity to spatial distortions and inability to account for perceptual nuances make it inadequate for tasks where temporal coherence and viewer experience are vital, as noted in studies on GANs and video quality evaluation frameworks [1].

An advancement in this metric landscape is the introduction of Fréchet Video Distance (FVD), which assesses the diversity and quality of video generation. FVD builds on the principles of Fréchet Inception Distance (FID) used in static images to gauge the statistical distance between feature distributions of real and generated video clips. Specifically designed for video data, FVD incorporates temporal coherence by taking into account multiple frames, thus aligning more closely with human judgment of quality. This correlation is supported by empirical evidence establishing FVD's robustness as a metric for emerging deep generative models of video [41].

Multimodal assessments introduce another layer by integrating audio-visual components into quality measures. Such metrics evaluate not only the visual fidelity but also the synchronization between audio and video elements, crucial for narrative coherence as observed in human feedback systems. This approach is especially valuable in generative tasks within frameworks such as video GANs and neural networks aimed at producing cohesive storytelling [33].

Human-centric evaluation metrics such as Mean Opinion Score (MOS) and Quality of Experience (QoE) have emerged in tandem with advancements in subjective assessments. MOS captures an aggregate rating from viewers directly influenced by their experiences, while QoE encompasses broader user satisfaction parameters. These evaluations resonate particularly well in applications that require viewer engagement, such as entertainment and marketing, where human perception significantly impacts perceived quality [2].

Despite their advantages, subjective metrics can introduce variability due to personal biases and contextual interpretation. Studies have highlighted that these biases may not align perfectly with objective metrics, creating a dichotomy that challenges researchers to strike a balance between them. Continuous advancements in crowdsourcing frameworks that collect viewer feedback represent an evolving response to this challenge, allowing for real-time adjustments and refinements in generative processes [5].

Emerging trends indicate a shift towards integrating advanced machine learning techniques to create adaptive quality assessment models that learn from human feedback. By employing reinforcement learning from human feedback (RLHF), models can iteratively improve their outputs based on detailed human evaluative data, ensuring a closer alignment with user preferences [3]. This paves the way for more nuanced evaluation systems capable of dynamically adapting to diverse feedback.

As video generation technologies advance, future directions will likely focus on hybrid metrics that incorporate both subjective and objective data, enhancing the reliability of evaluations. Additionally, the development of benchmarks like AIGCBench aims to provide standardized frameworks for comprehensive assessments, facilitating the comparison of various generative models across multiple tasks [42]. Such endeavors will be critical in addressing the ongoing challenge of harmonizing automated evaluations with qualitative human judgments in the rapidly evolving landscape of video generation.

4.2 Role of Human Feedback in Evaluation

The integration of human feedback into evaluation metrics represents a transformative shift from traditional automated assessments toward more nuanced methodologies that incorporate human insights. This section critically examines the role of human feedback in evaluating video generation outputs, emphasizing the comparative advantages of human-in-the-loop strategies as well as the inherent challenges they present.

Historically, video evaluation metrics relied heavily on objective measures such as Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM). While computationally efficient, these methods often fail to account for subjective aspects of quality that resonate with viewers, overlooking critical components like storytelling coherence and emotional engagement—factors that human evaluators naturally incorporate into their assessments [43]. As a response to these limitations, researchers have begun implementing approaches that blend traditional metrics with human feedback, leading to a more holistic evaluation paradigm.

Human feedback can be classified into direct assessments, where users rate videos based on specific criteria, and implicit feedback, derived from user interactions,

such as engagement rates or viewing durations. This dual approach facilitates comprehensive evaluation frameworks that balance human insights with quantitative data. For instance, incorporating user ratings into a reinforcement learning (RL) model can enhance video generation outcomes by creating a feedback loop in which human feedback influences future outputs through reinforced learning signals [4]. However, the challenge remains in accurately translating qualitative evaluations into actionable quantitative metrics that models can adapt.

A significant innovation in utilizing human feedback is the concept of reinforcement learning from human feedback (RLHF), which emphasizes aligning models with user preferences through preference models derived from human judgments. RLHF frameworks have shown considerable effectiveness in tasks requiring nuanced understanding, such as language and video generation, demonstrating significant improvements when leveraging direct human feedback compared to conventional RL techniques [44]. This suggests the potential for more adaptive systems capable of evolving according to shifting human needs and preferences.

Despite the advantages, employing human feedback is not without its drawbacks. One critical issue relates to the variability and potential bias inherent in human evaluations. Judgments can be influenced by individual experiences, cultural backgrounds, and contextual factors, introducing inconsistencies in feedback [11]. To mitigate this, many researchers advocate using aggregated user feedback across diverse demographics to establish a more representative evaluation standard [6].

Emerging trends indicate a growing interest in hybrid models that combine automated metrics with human feedback, a concept aptly termed 'post-hoc feedback alignment.' Techniques proposed in the literature, which utilize fine-grained feedback alongside traditional metrics, aim to bridge the gap between subjective human evaluations and objective computational assessments. This approach can be formalized through reward modeling techniques that adapt based on user feedback specifics while maintaining alignment with general automated evaluations [8].

In summary, the integration of human feedback into the evaluation process signifies a critical evolution in enhancing the quality of video generation outputs. To advance generative models, it is essential to deepen our understanding of effectively soliciting, interpreting, and acting upon human feedback. Future research should focus on developing statistically robust methods to analyze feedback variability, alongside creating comprehensive feedback collection platforms that ensure inclusivity and representativeness of user perspectives. This collaborative model of evaluation will foster the creation of generative systems better aligned with human expectations, enhancing the overall user experience in video generation applications.

4.3 Balancing Subjective and Objective Measures

Balancing subjective human feedback with objective quality metrics is paramount in developing robust evaluation frameworks for video generation systems. Human perception is inherently nuanced, making subjective assessments indispensable for capturing aspects of video quality that automated processes often overlook. Conversely, objective metrics provide repeatability and scalability, essential for assessing large datasets efficiently. A harmonized approach that effectively integrates both types of feedback can lead to a more comprehensive evaluation strategy, thereby enhancing the alignment of generated video content with user expectations.

One primary challenge lies in the disparity between subjective experiences and objective measurements. For instance, metrics such as Fréchet Video Distance (FVD) and Inception Score (IS) often quantify dimensions like visual fidelity and temporal coherence; however, they may not encapsulate emotional engagement or narrative coherence, which are critical in human-centered evaluations [41]. Traditional metrics can misrepresent the richness of human assessments, leading to potential discrepancies between what is produced by generative models and what resonates with audiences. Research has highlighted instances where subjective scores diverged significantly from those given by automated metrics, revealing a disconnect that can hinder technological progress [15].

Emerging models for integrating subjective and objective metrics often adopt a mixed-methods approach. For example, the VideoScore framework utilizes large-scale human feedback data to align automatic evaluations with human perceptions, achieving a correlation of 0.771 with subjective assessments [21]. This illustrates the potential of using reinforced learning paradigms that adapt model assessments based on human-preference scores, thus refining the evaluation landscape.

Certain techniques focus on calibrating automated outputs using human feedback. By employing reinforcement learning from human feedback (RLHF), for instance, models can be tuned to prioritize attributes deemed important by users, thereby aligning more closely with subjective evaluations without sacrificing the scalability of objective measures [5]. This dual-layered evaluation creates robust feedback loops that improve model performance iteratively.

Technical challenges also surface when attempting to balance these dimensions. On the one hand, subjective evaluations are labor-intensive and costly, particularly when high volumes of data are involved. On the other hand, reliance solely on automated metrics can lead to the oversight of crucial descriptive qualitative traits, such as originality or emotional resonance, which are often paramount in creative tasks [45].

Another emerging trend is the development of hierarchical evaluation systems that disaggregate video quality into specific attributes, such as motion smoothness, thematic alignment, and visual coherence. These systems allow evaluators to provide fine-grained feedback on various aspects, enabling models to develop responses tailored to user-defined quality dimensions [46]. As such, moving forward, a potential research avenue could consider building comprehensive multi-dimensional models that not only aggregate subjective feedback but also incorporate weighted scoring for diverse qualitative factors.

Practical implications of these methodologies suggest a deeper collaboration between human and machine interaction in creative domains. Recognizing the role of human preference could inspire the design of user-centric interfaces

that prioritize user experience alongside traditional model training metrics, ensuring models evolve in tandem with audience expectations [43]. This synthesis points towards a future where evaluations are not only about achieving technical excellence but also fostering genuine user engagement and satisfaction.

In conclusion, achieving a meaningful balance between subjective and objective evaluation methods is critical for advancing video generation technologies. Future research should continue to explore innovative frameworks that decrease reliance on singular measures by promoting multi-faceted strategies capable of adapting to evolving human preferences, ultimately leading to more sophisticated and engaging video content.

4.4 Evaluation Frameworks and Methodologies

Structured frameworks for evaluating generated videos play a crucial role in blending human feedback with automated assessments, providing both qualitative insights and quantitative performance metrics. These frameworks aim to create a holistic evaluation landscape that accurately captures the complexities inherent in video generation. An effective evaluation framework typically incorporates multiple components, including human evaluations, objective metric assessments, and feedback loops that facilitate continuous learning and adaptation in generative models.

One prominent approach leverages the multi-metric evaluation paradigm, combining human annotator scores with automated metrics to achieve a comprehensive understanding of generated video quality. Recent works have emphasized the importance of aligning automated metrics with human perception by utilizing human preference datasets to guide evaluation standards [46]. This multi-faceted approach mitigates the limitations associated with isolated metric evaluations, enabling a nuanced understanding of how generated content resonates with human viewers while complementing traditional quality measures such as Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) [27]. The flexibility to incorporate diverse metrics significantly enriches the evaluation process.

However, challenges persist in balancing automation with human evaluations. While automated metrics offer advantages in scalability, they often fall short in capturing the creative nuances that characterize human-generated content [47]. Additionally, the reliability of subjective human ratings can vary based on individual preferences or biases, potentially leading to inconsistencies in evaluation outcomes. Therefore, it becomes essential to synchronize qualitative assessments with robust statistical analyses, such as correlating user satisfaction scores with traditional metrics through techniques like Spearman rank correlation [27]. Such analytical methods can uncover trends that enhance our understanding of how various attributes contribute to overall user enjoyment and engagement.

Emerging frameworks increasingly lean towards the integration of dynamic evaluation methods that account for temporal factors in video quality, such as the consistency of motion and scene transitions. The DEVIL framework, for instance, introduces a series of dynamics-based scores that capture the fluidity and interactivity of generated videos, paving the way for a more comprehensive assessment matrix that extends beyond visual fidelity and content relevance [48]. By addressing the dynamics of video content, frameworks like DEVIL are moving towards a more contextualized evaluation that can cater to the evolving demands of both creators and audiences.

The proliferation of large-scale datasets, such as VideoFeedback, enables the training of sophisticated models that simulate fine-grained human feedback. This incorporation of extensive empirical data allows for the formulation of personalized evaluation metrics tailored to specific user needs and preferences. The evolution towards human-aligned benchmarks signifies a shift in focus where understanding user experience takes precedence over mere technical performance metrics [49].

As the field advances, there is a noticeable trend toward leveraging advancements in machine learning and artificial intelligence for automated feedback analysis. Evaluative frameworks may integrate advanced models capable of generating synthetic evaluations based on historical human feedback, enabling generative systems to adaptively fine-tune their outputs [50].

In conclusion, the landscape of video generation evaluation is on the verge of transformative innovation, driven by the synthesis of human feedback and automated methodologies. Future research directions should concentrate on refining these frameworks to enhance their robustness, addressing existing challenges related to subjective biases in human evaluations while capitalizing on the potential of automated systems. The ongoing dialogue between evaluative frameworks and generative modeling holds significant promise for advancing the field, fostering a deeper understanding of how to create compelling, engaging video content through iterative, feedback-informed processes.

4.5 Future Directions in Video Evaluation

The evaluation of video generation models is rapidly evolving, particularly as it intersects with human feedback integration mechanisms. Emerging trends indicate a transition towards more nuanced, multi-dimensional assessment frameworks that not only prioritize automated scoring but also capture the qualitative aspects of human viewer engagement. Current metrics, while effective in certain contexts, often fail to align adequately with human perceptions of video quality, motivating a need for more sophisticated evaluative methods. The integration of human feedback into evaluation metrics is fundamental to this pursuit, as it provides both diversity in feedback types and richness in contextual understanding.

A significant direction for future research is the refinement of fine-grained feedback systems, which allow for more precise insights into viewer preferences. Fine-grained feedback encompasses detailed evaluations that dissect specific attributes of video outputs, such as storytelling coherence, visual fidelity, and emotional resonance. For instance, the inclusion of metrics that assess the impact of narrative pacing or character development aligns closely with viewers' subjective experiences, as shown in recent works on enhancing video coherence through detailed viewer interaction studies [51]. Furthermore, integrating automated

assessments, such as VideoScore—which leverages a dataset of multi-aspect scores from human judges—facilitates the benchmarking of video quality by simulating human feedback more effectively than existing metrics [21].

Simultaneously, the avenue for dynamically adaptive evaluation metrics that evolve based on continuous human feedback is gaining traction. Such metrics can utilize reinforcement learning paradigms that adaptively recalibrate evaluation scores to reflect the evolving landscape of viewer preferences. This approach is supported by studies indicating that feedback loops can significantly enhance model alignment with human judgement, which is critical for addressing the disparities found in traditional metrics [52]. This method stands in contrast to static evaluation techniques, thus promoting a more responsive and iterative evaluation process.

Another emerging trend focuses on the ethical implications of human feedback mechanisms in video evaluation. As crowd-sourced data collection becomes more prevalent, attention must be directed toward establishing protocols that ensure equity and mitigate biases inherent in feedback from varied demographic groups. This emphasizes the need for careful dataset design and the incorporation of diverse viewpoints to guide evaluative metrics, thereby enhancing the representational fairness of generated content [53].

Moreover, the future landscape of video evaluation metrics will benefit from the integration of psychological and emotional analysis tools that assess user engagement beyond surface-level interactions. Techniques such as sentiment analysis could be employed to gauge viewer reactions in real-time, integrating systems that measure emotional feedback during dynamic video viewing experiences—an intersection of viewer behavior analytics and video generation that fosters deeper insights into content effectiveness [54].

As advancements in high-dimensional space representation and reduced memory consumption techniques are being developed [55], the ability to effectively model and measure spatial-temporal dynamics of video content is paramount. For instance, modeling video generation through state-space approaches demonstrates promising efficiency in accommodating longer video sequences while maintaining quality [55]. This technical evolution dovetails with refined human evaluative responses, recognizing that audiences perceive quality not merely through isolated frames but rather through interconnected narratives conveyed over time.

In conclusion, the future of video evaluation metrics lies in creating integrated evaluation systems that capitalize on both automated and human feedback mechanisms. By utilizing fine-grained assessments, dynamic adaptation processes, and ethical frameworks, we can establish more robust evaluation paradigms that are reflective of human preferences, ultimately driving the convergence of technology with human-centric video generation practices. The anticipation of these developments offers an exciting pathway for future exploration, compelling researchers and practitioners to rethink not just how we measure video quality, but how we ensure it resonates meaningfully with its audience.

5 Applications and Case Studies

5.1 Enhancing User Engagement in Entertainment

Human feedback has emerged as a transformative component in enhancing video generation within the entertainment industry, particularly in films and animations. The integration of user interactions into creative processes not only fosters a more engaging viewing experience but also enables filmmakers and animators to better align their outputs with audience preferences. This subsection will delve into various approaches that leverage human feedback to enhance audience engagement, providing a comparative analysis of their effectiveness, strengths, and limitations.

One prominent method employed in the entertainment domain is the use of interactive feedback mechanisms within animated films. Here, audience input is solicited to influence narrative arcs, character development, and even pacing. This is exemplified in participatory films where viewers can vote on story elements, effectively creating a customized viewing experience. Studies have shown that this audience-driven model results in more relatable and compelling narratives, as viewer preferences directly shape the story progression, leading to increased emotional investment ([4]).

Streaming platforms have also harnessed human feedback to tailor content dynamically. For instance, algorithms adjust video styles, pacing, and even thematic elements based on real-time audience interactions. By analyzing viewer data—such as watch time, skips, and engagement metrics—the platforms can quickly adapt their offerings to match shifting audience preferences. Such adaptive systems leverage reinforcement learning techniques, allowing for a continuous feedback loop that informs real-time adjustments ([42] [47]). This model not only improves user satisfaction but also enhances retention rates, presenting a clear competitive advantage in an increasingly fragmented entertainment market.

Community-driven feedback systems represent another innovative avenue for engagement. Filmmakers and production studios increasingly utilize platforms that solicit viewer opinions during various stages of production. This collaborative environment enriches the creative process, allowing creators to gather diverse perspectives that can lead to more nuanced and culturally relevant storytelling. Notably, films like “Crowdsourced Cinema” exemplify this approach by integrating feedback from prospective audiences into production decisions, thus cultivating a sense of ownership among viewers ([1]). However, while these systems enhance engagement, they also pose challenges regarding the representation of diverse voices and the potential marginalization of minority perspectives within mainstream narratives.

In evaluating the effectiveness of these approaches, it is essential to consider their strengths and trade-offs. Interactive feedback mechanisms in animation promote heightened viewer engagement and aim to cultivate community ties but can also dilute artistic intent if poorly managed. Streaming platforms’ reliance on data-driven algorithms enables rapid adjustments; yet, it risks homogenizing content to appeal to mainstream preferences, potentially stifling diversity in creative expression. Community-driven systems galvanize

collective input, enriching content relevance but can sometimes lead to conflicts between audience desires and artistic originality, as creators navigate the balance between market demands and personal vision.

Emerging trends indicate a growing sophistication in how human feedback is integrated into entertainment-driven video generation. The use of AI-driven predictive analytics, for instance, allows for proactive content adjustment based on historical feedback data, enabling creators to anticipate audience reactions ([3]). Moreover, future advancements may see the incorporation of wearables that capture emotional responses during playback, facilitating a transformative feedback mechanism that resonates with viewer experiences on a visceral level.

In summary, integrating human feedback into video generation within the entertainment industry presents significant opportunities for enhancing audience engagement. While various approaches demonstrate differing strengths and limitations, the continued evolution of these systems promises to cultivate more dynamic and responsive content creation practices that can address the nuanced demands of contemporary audiences. As filmmakers and technologists explore innovative direction, the interplay between art and audience engagement will undoubtedly redefine storytelling in the digital age.

5.2 Personalized Learning Experiences in Education

The integration of human feedback into educational video generation has emerged as a pivotal strategy for crafting personalized learning experiences tailored to the unique needs of students. By leveraging adaptive content techniques, educational platforms can synthesize real-time feedback from learners, allowing for the modification and optimization of video materials that foster engagement and enhance comprehension. Various case studies underscore the effectiveness of these approaches, illustrating their potential to fundamentally transform traditional educational models.

One prominent application of personalized learning through human feedback can be observed in interactive educational environments that utilize real-time assessments to dynamically tailor video content. Notably, platforms like Khan Academy have experimented with feedback mechanisms allowing learners to indicate their levels of understanding after consuming video lessons. By analyzing these inputs, the system can adjust the subsequent content presented, altering difficulty levels or revisiting foundational concepts as necessary. This approach supports the premise that individualized feedback within educational frameworks can lead to markedly improved learning outcomes, aligning with findings from studies on reinforcement learning from human feedback [5].

Furthermore, the deployment of AI-generated instructional materials has produced adaptable video content that resonates with students' interests and learning styles. Recent studies highlight how personalized podcasts, formulated based on user preferences, have significantly enhanced engagement rates among learners. By enabling AI to curate educational content that reflects individual trajectories, these platforms foster deeper connections with the material [2]. This trend towards AI-generated educational media is part of a larger movement focusing on data-driven personalization in education, which has been shown to improve overall student performance and satisfaction metrics.

However, the integration of human feedback into educational video generation also presents several challenges. The quality and reliability of feedback can vary significantly among users, leading to inconsistencies in content adaptation. Identifying the most effective feedback mechanisms remains an important area of exploration, requiring systems to differentiate between constructive critiques and outlier responses [56]. Moreover, fine-tuning content based on intermittent human feedback necessitates robust algorithms capable of processing diverse input types while maintaining efficiency in learning.

Another notable trend is the increasing implementation of community-driven feedback systems within educational contexts, where learner cohorts contribute collectively to the development of instructional content. This collaborative approach has proven beneficial for enhancing both the relevance and richness of the generated educational material. By synthesizing feedback from large groups of users, platforms can create a dynamic repository of instructional resources that continually evolves, reflecting the collaborative design methodologies observed in creative AI and interactive media [57].

Despite these advancements, significant hurdles remain in ensuring that educational tools are accessible and equitable across diverse learner populations. The challenge lies in balancing personalized content with universal design principles, guaranteeing that all students can benefit from adaptive learning technologies. As educational stakeholders continue to explore the integration of human feedback mechanisms, efforts must focus on enhancing inclusivity and addressing the unique challenges faced by marginalized groups [6].

In conclusion, the landscape of personalized learning experiences facilitated by human feedback in educational video generation is both promising and complex. The synthesis of human inputs into adaptive educational platforms holds the potential to refine content delivery, thereby improving learner engagement and comprehension. Continued research into feedback reliability, content adaptability, and scalability will be essential for overcoming existing challenges and enriching the educational experience through technology. As this field evolves, innovative methodologies and collaborative frameworks will be crucial in harnessing the full potential of human feedback in educational contexts, paving the way for more effective personalized learning solutions.

5.3 Marketing Strategies Leveraging Audience Feedback

The application of human feedback in generating promotional videos for marketing campaigns has emerged as a game-changing strategy for optimizing content relevance and enhancing consumer engagement. By examining audience feedback mechanisms, marketers can tailor video content to align closely with audience preferences, thereby maximizing the impact of advertising efforts. Analyzing existing methodologies reveals diverse approaches that leverage

consumer input, each with inherent strengths, limitations, and contextual trade-offs that shape their effectiveness.

One prevalent approach incorporates automated tools that refine video content based on audience feedback. For instance, machine learning algorithms can analyze sentiment from audience reactions—measured through engagement metrics such as likes, shares, and comments—to iteratively enhance marketing strategies. This adaptive feedback loop allows brands to adjust their messaging dynamically, ensuring alignment with consumer expectations. Approaches utilizing generative adversarial networks (GANs) can synthesize content that resonates stronger with audiences when trained on such feedback data [1].

In contrast, engaging consumers directly through interactive platforms yields rich qualitative insights that can inform video narratives and aesthetics. Techniques such as focus groups and audience testing sessions foster direct involvement, allowing audiences to express preferences and critiques that can guide content creation. For example, the integration of human evaluative input into hybrid systems—combining GANs with audience feedback—has been shown to improve narrative coherence and viewer engagement, suggesting a pathway for further enhancing promotional videos [58].

Strengths of these approaches include their capacity to ensure greater alignment with emerging consumer trends, as well as their potential to elicit emotional connections with viewers. However, the efficacy of these strategies is contingent upon the systematic analysis of feedback quality and the management of biases. For example, implicit feedback mechanisms—like analyzing viewer drop-off rates or interaction patterns—may not capture the context of audience sentiment effectively, leading to misleading interpretations of consumer preferences [16]. Moreover, an overreliance on automated feedback systems can obscure the nuances required for genuine connection, as real-time emotional responses may differ from engagement metrics alone.

Emerging trends in leveraging audience feedback are characterized by the increased utilization of A/B testing frameworks that facilitate iterative video modifications. Using methods that analyze side-by-side viewer reactions to differing video elements enables more nuanced insights into audience preferences, allowing for targeted content adjustments before large-scale releases. This approach has been enhanced by insights into multi-modal models capable of processing both visual and auditory feedback simultaneously, resulting in more holistic content customization [15].

Nevertheless, challenges remain in striking a balance between automated feedback systems and human judgment, particularly in evaluating the subjective qualities of generated content. The recent advent of comprehensive evaluation frameworks, such as the VBench benchmark suite, highlights the need for consistent standards that effectively merge qualitative human insights with quantitative data analysis [46].

Future directions in marketing strategies that harness audience feedback must focus on scalable feedback collection methods that do not impede the user experience. The development of lightweight, real-time feedback tools, combined with explainable AI techniques, could enhance engagement without overwhelming consumers [59]. Additionally, ethical considerations surrounding data privacy must remain at the forefront as brands seek to harvest audience insights. Establishing clear communication about how feedback data is utilized could foster trust and encourage participation, amplifying the power of audience-driven content refinement.

In conclusion, the integration of audience feedback into promotional video generation offers a robust framework for enhancing marketing efficacy. By combining automated tools with human-centered approaches, marketers can create engaging, relevant content that resonates with viewers, underpinning the future of adaptive consumer interactions in advertising.

5.4 Cross-Domain Insights from Creative Feedback Platforms

Cross-domain applications of human feedback systems in video generation exemplify a collaborative synergy that fosters innovation across various creative platforms. By integrating human feedback mechanisms, the quality and relevance of generated content can be significantly enhanced in disparate fields such as gaming, education, and marketing. For instance, in gaming environments, feedback systems enable iterative content creation, where community-driven insights play a crucial role in shaping narrative development and character design. This methodology closely resembles the crowdsourced evaluation techniques explored in works like [35], whereby user input directly influences adaptable learning curves and refines the gaming experience.

In educational contexts, the incorporation of real-time feedback systems has proven successful in personalizing learning pathways. Adaptive video generation tools can dynamically adjust content to match students' proficiency levels, leveraging feedback loops to foster greater engagement. This adaptability mirrors findings in research, such as [33], where real-time evaluations effectively bolster learner involvement, showcasing the versatile potential of integrating creative feedback systems across educational platforms.

Moreover, in the advertising sector, utilizing human feedback for optimizing promotional video content has practical implications. Marketing campaigns have increasingly adopted iterative feedback processes, where consumer reactions to initial video drafts inform subsequent revisions. This strategy not only enhances viewer engagement but also boosts conversion rates by aligning content closely with audience expectations. The efficacy of this method is evident in studies like [60], which emphasizes the importance of integrating consumer insights in content curation.

The technical nuances revealed through cross-domain interactions yield critical insights into the strengths and limitations intrinsic to each field. While reliance on human feedback enriches the contextual depth that automated systems may lack, challenges such as balancing the volume of feedback and ensuring data quality persist across various platforms. For example, rapid adaptation to user insights in gaming may lead to content inconsistencies, as noted in [15]. Conversely, the slower approach often adopted in educational settings may facilitate deeper content refinement but risks limiting responsiveness to real-time feedback.

Emerging trends illustrate a shift toward more elaborate interaction strategies, including the use of multi-modal

feedback systems that synthesize user input from diverse sources. As explored in [61], the integration of video, text, and interactive elements points to a future where cross-domain systems can comprehensively gather and analyze feedback, enhancing user experience. This adaptability represents a paradigm shift, facilitating the evolution of narrative structures and character arcs in alignment with audience preferences.

As innovations progress, ensuring the ethical implementation of feedback systems takes on increasing significance. Addressing biases inherent in user feedback—particularly in diverse demographic settings—is critical. Techniques outlined in [31] examine the extraction of insights from community dynamics, which are instrumental in achieving equitable video generation outcomes. By designing systems that acknowledge cultural nuances within feedback mechanisms, developers can ensure inclusivity, thereby expanding the reach and relatability of generated content.

In conclusion, the ongoing synergy among human feedback platforms across diverse domains not only drives innovation in video generation but also highlights an intricate balance of responsiveness, cultural relevance, and ethical considerations. The future is likely to witness an escalation of adaptive approaches that refine video generation through multi-dimensional feedback, enriching the user experience while navigating the complexities of audience interaction and engagement. Such endeavors promise to redefine creative frameworks and nurture richer, more meaningful content generation avenues across various applications.

5.5 Future Directions and Innovations in Video Generation

As the landscape of video generation continues to evolve, the integration of human feedback stands out as a critical factor for future advancements. Emerging trends indicate a converging path towards more interactive, real-time, and adaptive video generation technologies that harness the nuances of human preferences. One notable direction is the enhancement of reinforcement learning frameworks, where algorithms can adapt to human critiques in real time. Techniques such as $RLHF$ (Reinforcement Learning from Human Feedback) have already demonstrated potential in aligning AI behaviors with user expectations by utilizing sparse and non-expert feedback effectively [62].

A pivotal innovation in this space is the exploration of multimodal feedback systems that encompass various forms of human interaction, extending the applicability of feedback beyond mere ratings to include qualitative inputs, emotional data, and contextual signals. For instance, advancements in wearable technology could enable applications that collect physiological indicators while users consume video content, thus facilitating a more profound understanding of user engagement and preferences in content generation [51]. This could lead to the development of dynamic systems that adjust video outputs in real time based on emotional reactions, thus increasing viewer satisfaction and engagement.

Moreover, the refinement of automated evaluation metrics remains a salient research focus. Current tools often fail to resonate with human evaluative criteria, as evidenced by the limitations of traditional metrics in correlating well with user satisfaction [46]. The development of sophisticated metrics, like VideoScore, which correlate with human ratings and account for fine-grained aspects of user feedback, exemplifies the shift towards creating benchmarks that reflect a more holistic understanding of video quality [21]. By utilizing datasets such as VideoFeedback, which offers multifaceted human scoring, future models could be fine-tuned to better reflect viewer preferences and biases, leading to more broadly acceptable outputs.

Additionally, the approach of integrating AI feedback with human input could prove transformative. RLAIF (Reinforcement Learning from AI Feedback) emerges as a promising avenue, offering a method to streamline the feedback collection process while maintaining the alignment of generated outputs with user intent [40]. The effectiveness of such methodologies lies in their ability to provide rapid, inexpensive feedback that can scale to accommodate large populations, significantly enhancing the training datasets available for video generation models [22].

Challenges persist, primarily regarding biases introduced through human feedback and the variability in user preferences. Addressing these limitations requires ongoing research into system designs that ensure diversity and inclusivity of feedback sources. Approaches utilizing crowdsourced feedback provide a promising direction but necessitate careful consideration of the representational bias that may emerge from these datasets [2]. Furthermore, the establishment of ethical guidelines for data collection and user privacy will be paramount in ensuring that feedback systems do not inadvertently marginalize cohort voices.

Innovatively, hybrid models combining generative techniques with feedback mechanisms are encouraging further exploration into creative video outputs. For instance, models that utilize adversarial training, supported by human feedback, can enhance creative outputs by allowing models to explore deviations from norm styles [63]. These systems can facilitate not only improved realism but also creativity by understanding subjective user experiences in video narratives.

In conclusion, the integration of human feedback into video generation is set to redefine the capabilities and performance of modern generative models. By embracing advancements in real-time feedback systems, developing robust evaluation metrics, and addressing ethical considerations, future innovations are poised to deliver superior video content that resonates profoundly with audience preferences. The resulting advancements are expected to shape not only the technological landscape but also the cultural narratives that these generated videos will embody.

6 Challenges and Future Research Directions

6.1 Scalability of Human Feedback Integration

Integrating human feedback into video generation systems at scale introduces several complex challenges associated with diverse user populations and extensive datasets. The efficacy of these systems largely depends on their ability to accurately interpret and process varied inputs from a

broad user base, which can include both expert and non-expert feedback. As these models evolve, the irregularities and biases inherent in user-generated data become increasingly prominent, necessitating sophisticated methods for feedback normalization and utilization.

One major challenge is the adaptability of feedback mechanisms in handling inputs that vary significantly in quality, specificity, and relevance. Systems designed to integrate human feedback must incorporate techniques that enable them to weigh the feedback based on its source, allowing for a more coherent integration process. For example, approaches such as Reinforcement Learning from Human Feedback (RLHF) have shown promise in tuning models to better align with human preferences by making use of feedback from users to refine outputs dynamically [5]. However, this method also raises concerns about scalability due to the resource-intensive nature of collecting and processing extensive human feedback.

Emerging trends highlight the necessity for scalable feedback integration systems that can process implicit user feedback – such as viewing time and click-through rates – alongside explicit evaluations like ratings or comments. Implicit feedback systems analyze user interactions and engagement metrics to deduce preferences without requiring direct input, as described in recent works [50]. These systems can adapt quickly to large datasets, providing real-time updates that can dramatically enhance video generation quality, yet they also face challenges related to data bias and accuracy in inferred preferences.

Another significant challenge is integrating diverse feedback within the same framework. Human feedback varies not only in quality but also in modality and context. Systems must evolve to accommodate this diversity without compromising the coherence and reliability of the generated outputs. Techniques leveraging multi-modal data integration are developing, with models that can draw on both visual and textual inputs to better capture user feedback across different contexts [64]. This dual-input approach emphasizes how integrating various forms of feedback can lead to more refined and relevant outputs.

Technological advancements in machine learning frameworks are pivotal to addressing these scalability issues. For instance, researchers are exploring using hierarchical models that allow for the processing of feedback at multiple granularity levels [65]. Such models can prioritize high-quality feedback and normalize less reliable inputs effectively. They balance immediate responsiveness to feedback with long-term model adaptation that supports iterative improvement over time.

Looking forward, the implementation of comprehensive evaluation metrics that align human feedback with model performance across different dimensions of video generation quality is essential. Standardized benchmarks, such as VBench, can facilitate a more coherent assessment framework that not only accounts for technical metrics but also incorporates qualitative human judgments [46]. This holistic evaluation approach could bridge the gap between user feedback and model development, encouraging models to learn from user experiences over time.

In conclusion, addressing the scalability of human feedback integration in video generation requires innovative methodologies that account for the diversity and complexity of user feedback. Balancing explicit and implicit feedback, leveraging multi-modal integration techniques, and developing robust evaluation frameworks are critical next steps in evolving these systems. Future research should focus on refining these approaches to foster a dynamic and responsive video generation environment that aligns closely with user expectations and preferences.

6.2 Ethical Considerations in Data Collection

The collection of human feedback for video generation, while invaluable for enhancing model performance, poses a range of ethical concerns, particularly regarding user privacy, consent, and potential biases. Establishing a transparent framework for data collection is essential to address these ethical dimensions thoughtfully, mitigating risks associated with participant engagement and fostering trust in AI systems.

At the heart of ethical data collection lies the principle of informed consent. Participants providing feedback should clearly understand the intended use of their data, who will have access to it, and the measures in place to safeguard their privacy. This necessitates effective communication about the objectives of data collection and the technical practices employed to protect participant information. For example, human-in-the-loop systems can include explicit clauses that define how user data will be utilized, ensuring participants can withdraw their contributions without penalties. This concern is particularly relevant in scenarios where feedback data is rich in personal sentiment, making inadequate anonymization a notable risk if the information is mishandled [33].

Another significant ethical concern involves the potential for bias in the data collected. Human feedback is inherently subjective, influenced by users' personal experiences and sociocultural contexts. Consequently, algorithms utilizing this feedback as a learning signal risk perpetuating existing biases, especially when feedback predominantly comes from homogeneous groups. To address this, it is crucial to implement protocols that foster diversity within the participant pool to capture a broader array of perspectives. Techniques proposed in the literature suggest incorporating a wider range of feedback types—such as employing ordinal alongside cardinal feedback—to help mitigate the biases associated with relying on single-feedback scenarios [51].

Emerging practices in ethical data collection leverage advanced computational techniques to further these considerations. For instance, frameworks that integrate reinforcement learning from human feedback (RLHF) can directly encode fairness criteria into the learning process. By incorporating fairness constraints during model training, systems can learn not only to optimize performance but also to adhere to ethical standards regarding the feedback utilized [2]. Additionally, employing implicit feedback mechanisms—where user interactions with content, like engagement metrics, are analyzed—can optimize the feedback process while reducing reliance on explicit subjectivity, which can introduce ethical complexities. However, this approach also raises concerns about data accuracy and the risk of misinterpreting user engagement as a positive endorsement [66].

Furthermore, the ethical landscape is complicated by the choices surrounding data management post-collection. Protocols for data storage, sharing, and deletion must respect user autonomy and minimize potential harm. Under the General Data Protection Regulation (GDPR) and similar legislative frameworks, organizations are mandated to ensure that users have rights concerning their data, including access, rectification, and erasure. Neglecting to adhere to these ethical guidelines can lead to significant legal ramifications and erode public trust in technology [43].

As the field progresses, future research must prioritize the establishment of mechanisms that enhance ethical standards in data collection for human feedback in AI systems. This includes the development of automated systems that integrate ethical oversight throughout the data lifecycle, proactively addressing potential misuses or biases that may arise during AI training. Promising directions involve deploying ethical auditing tools capable of evaluating how feedback mechanisms align with established ethical guidelines throughout the development process [67].

In conclusion, ethically justifying the collection of human feedback for video generation necessitates a multifaceted approach that harmonizes participant rights with technological advancement. By prioritizing transparency, bias mitigation, and ongoing ethical scrutiny, researchers can develop models that not only perform effectively but also respect and value human contributions, setting a precedent for future AI systems that emphasize ethical integrity alongside efficacy.

6.3 Methodological Improvements for Feedback Efficiency

Enhancing the efficiency of integrating human feedback into video generation systems is critical to not only elevate the quality of the generated content but also optimize processing time and resource allocation. The persistent challenge lies in balancing the timeliness and accuracy of feedback mechanisms with the complexity of video generation models that necessitate intricate temporal relationships among frames. Emerging methodologies provide promising avenues for addressing these concerns.

One noteworthy approach is the employment of hybrid learning frameworks that seamlessly integrate user feedback with generative model outputs. For instance, affective feedback synthesis techniques can adapt video content in real-time based on users' emotional responses, creating a more intuitive interaction model. Such an interactive mechanism allows for instantaneous adjustments, ensuring that the system learns user preferences dynamically, thus enhancing the relevancy of generated videos in near real-time. The effectiveness of this feedback-focused adaptation highlights the importance of immediate engagement in learning systems, as underscored in recent studies exploring human feedback-driven models [8], [17].

Moreover, the implementation of reinforcement learning strategies presents a robust methodology for optimizing feedback efficiency. By utilizing approaches such as Reinforcement Learning from Human Feedback (RLHF), systems can refine their outputs based on user evaluations, generating a closed-loop learning environment. This process necessitates constructing efficient reward functions that correctly represent user preferences, which can be achieved using algorithms that aggregate feedback scores across multiple instances. For example, techniques like multi-dimensional preference scoring allow models to evaluate user choices along different criteria, improving feedback granularity and thus the quality of the generated videos [7], [68].

The trade-off between computation time and model fidelity can also be explored through multi-modal learning architectures that condition video generation not solely on textual descriptions but also incorporate visual inputs. This fusion enhances contextual understanding and reinforces user intent, thereby streamlining the generation process to yield more contextually relevant outputs. Additionally, conditioning frameworks like VideoMotion Customization utilize user-specified motion trajectories, which facilitate user control over nuanced aspects of the video generation without necessitating extensive retraining, thus preserving computational efficiency [14], [69].

Nevertheless, the integration of user feedback is not without challenges. One significant hurdle is the inherent variability of human feedback, which can be subjective and context-dependent. Establishing a standardization mechanism through extensive human-in-the-loop evaluations is essential to mitigate biases and improve the reliability of feedback loops. For example, developing comprehensive benchmarking datasets that encompass diverse user-generated inputs and preferences can enhance the representation of user needs within generative models [46], [47]. This commitment to refining human evaluative metrics will ensure that feedback incorporated into the systems is both accurate and aligned with common user expectations.

Future directions for research in this domain should focus on the development of transparent feedback methodologies that empower users to contribute to model training actively. Systems that allow users to specify constraints dynamically during the generation process will promote co-creative interactions, putting human intent at the forefront of video generation. Furthermore, leveraging advancements in automated feedback analysis tools can significantly reduce the processing duration of human inputs, thereby fostering a more agile generative process [13], [27].

In conclusion, methodological improvements for feedback efficiency in video generation hinge upon innovative learning frameworks, hybrid approaches to user engagement, standardization of evaluative metrics, and proactive user involvement. By concentrating on these areas, future research can create more responsive and satisfactory video generation systems that not only meet user expectations but evolve alongside them, translating feedback into actionable insights that drive consistent improvement in content quality.

6.4 Future Directions in Human-Informed Video Generation

Ongoing research is essential to uncover the vast potential of human feedback in innovating video generation methodologies and enhancing user experiences. Human-centric approaches to video generation stand to gain significantly by leveraging the diverse information provided through feedback mechanisms. The interplay between generative models

and user input yields rich opportunities for improvement in generative fidelity, content relevance, and emotional resonance with audiences.

Recent trends highlight the integration of more sophisticated human feedback systems in generative models, where techniques like reinforcement learning from human feedback (RLHF) [33] have demonstrated significant improvements in generating content that aligns with user expectations. By combining explicit feedback—such as ratings and critiques—with implicit behavioral metrics, these models can dynamically adjust outputs in real-time, rendering a more intuitive and interactive user experience. In contexts like interactive storytelling, continuous human input can shape narrative arcs and character behaviors, fostering deeper engagement—an approach evident in systems like Plan, Write, and Revise [70].

The concept of multi-modal training further expands this interaction, allowing models to accommodate various feedback forms while incorporating contextual understanding from both text and visual inputs [61]. This shift enhances video quality and addresses challenges in aligning complex user intentions with model outputs. Emerging frameworks that facilitate real-time collaborative input, such as DragNUWA [71], underscore the importance of developing seamless interfaces for user interaction.

However, notable trade-offs and limitations in these approaches must be acknowledged. For instance, while integrating RLHF can guide generative models closer to user desires, it may paradoxically lead to overfitting on specific preferences, thus reducing the exploratory capabilities of generative processes [72]. Furthermore, reliance on large datasets of human feedback can introduce noise and bias, complicating the fine-tuning process. Achieving a balance between user alignment and creative diversity remains a key challenge in the field.

Emerging trends in video generation are positioning the discipline towards greater personalization and adaptability. Advanced frameworks like VideoComposer and Control-A-Video demonstrate the utility of motion vectors and conditional inputs to fine-tune visual storytelling elements according to viewer engagement [27], [73]. These developments represent a significant step toward generative models that actively learn from users in a continuous feedback loop, evolving with their preferences over time. The potential for systems to conduct meaningful cross-modal interactions creates avenues for innovations in user-generated content.

Future directions should also involve developing robust evaluation metrics for assessing video generation quality alongside the integration of human feedback. As evidenced by initiatives like VideoScore, quantifying user satisfaction and aligning it with objective performance metrics is crucial for future model training [21]. Such frameworks should be expansive, allowing for layered evaluations that capture not only visual fidelity but also narrative coherence and viewer engagement.

In summary, the path forward for human-informed video generation is fraught with challenges but ripe with opportunities. The convergence of multi-modal feedback systems, advanced modeling techniques, and user-centric evaluation metrics will likely define future research trajectories, paving the way for generative models that are reactive and proactive in their interactions, consequently enhancing the user experience in video generation tasks. Addressing the intricacies inherent in this integration will solidify the foundational elements for creating rich, engaging, and personalized video narratives.

6.5 Comprehensive Evaluation Frameworks

The development of robust evaluation frameworks for assessing the integration of human feedback in video generation processes is crucial for advancing the field. Such frameworks must not only quantify the quality of generated videos but also consider the subjective nature of human preferences and their alignment with the content generation mechanisms. An effective evaluation framework should encompass both qualitative and quantitative measures, enabling comprehensive insights into the strengths and weaknesses of various generative models.

Current approaches to evaluating video generation primarily rely on traditional metrics such as Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM), which measure pixel-level fidelity. However, these metrics often fail to reflect human perception and engagement adequately. For instance, models like VideoScore propose more nuanced metrics that correlate well with human evaluations by capturing multi-dimensional aspects of video quality, such as temporal coherence and content relevance, which are critical for alignment with user expectations [21]. The limitations of existing evaluation methods highlight the need for multi-faceted criteria that go beyond mere pixel similarity to encompass elements like storytelling coherence, aesthetic appeal, and emotional engagement.

Recent advancements in reinforcement learning have led to frameworks that embed feedback directly into the generative process. The COnvergent Actor-Critic by Humans (COACH) framework emphasizes learning from critiques rather than predefined rewards, allowing models to adaptively refine outputs based on real-time human feedback [62]. However, as empirical evaluations show, the reliability and interpretability of these human feedback signals vary significantly, raising challenges in constructing benchmarks that can consistently predict user satisfaction across diverse contexts.

Moreover, longitudinal studies investigating how human feedback evolves over time provide valuable insights into the dynamic nature of viewer preferences. For example, innovative metrics developed in the context of video generation, such as those proposed in VBench, dissect generation quality into specific hierarchical dimensions, aligning closely with human assessments [46]. The incorporation of such insights into evaluation frameworks could enhance their robustness, allowing for more detailed diagnostics of generative model performance and guidance for iterative improvements.

An emerging trend within the evaluation of video generation models involves the integration of multi-modal data, which combines visual, textual, and audio feedback to form a holistic view of generated content. Techniques like those implemented in VideoControlNet leverage control signals from different modalities, which can be evaluated using feedback forms to ensure that the generated outputs adhere

to user-defined specifications [74]. This convergence of modalities presents opportunities for creating richer feedback loops that not only improve the quality of generated videos but also enhance user engagement by fostering greater interaction with the generative process.

Despite these advancements, challenges remain in balancing automated evaluation metrics with qualitative human assessments, particularly concerning biases that different demographic groups may exhibit in their feedback. Techniques for calibrating automated scores to better reflect human judgments have been proposed but require careful consideration of ethical implications concerning user data and privacy [53]. Future directions should aim for developing evaluation protocols that are not only systematic in their approach but also adaptable to various user groups, ensuring inclusivity in feedback collection.

In conclusion, advancing the field of video generation with integrated human feedback necessitates the development of comprehensive evaluation frameworks that address existing limitations. By focusing on multi-dimensional assessment criteria, enhancing the reliability of human feedback, and integrating multi-modal approaches, researchers can cultivate generative models that are not only effective in producing high-quality content but also finely attuned to human preferences. Continued exploration in this area is essential for fostering more engaging and contextually relevant video generation systems.

7 Conclusion

The integration of human feedback into video generation represents a pivotal advancement in the field, serving to bridge the gap between algorithmic efficiency and user-centric creativity. This survey elucidates various methodologies that harness human insights to refine video quality and align generated content with user expectations. Key frameworks such as reinforcement learning from human feedback (RLHF) have emerged as effective mechanisms, emphasizing the value of subjective evaluations over traditional, rigid reward signals. For instance, the use of models like TAMER and COACH allows for the adaptation of generative models through preference-driven approaches, illustrating the efficacy of utilizing direct human feedback in the training loop, as established by [2].

Comparatively, hybrid generative models incorporating human evaluations—e.g., generative adversarial networks (GANs) integrated with user feedback—have exhibited promising results. These systems enable the nuanced generation of high-quality videos by leveraging human evaluative signals to refine creativity and realism, showcasing a significant evolution in content generation paradigms. The Creative Adversarial Network (CAN) framework exemplifies this approach, as it not only integrates user feedback into the generation process but also actively encourages deviation from conventional norms to foster unique outcomes [1]. However, while these methods show substantial potential in enhancing output quality, they also unveil challenges, particularly in regards to the scalability and efficiency of feedback integration across diverse user populations.

Emerging trends indicate that interactive, user-guided video generation systems are gaining traction, offering real-time adjustments based on user feedback. These systems empower creators—allowing for dynamic participation during the generation process—which could fundamentally alter engagement strategies in media and entertainment [41]. The implications of these advancements extend across various applications, including education and advertising, where personalized experiences are increasingly prioritized. Yet, as the industry moves towards real-time human feedback interactions, critical challenges persist regarding the collection of reliable data and the management of biases inherent in crowd-sourced methods.

Furthermore, the development of benchmarks for evaluating the impact of human feedback on video quality, such as the proposed VBench and AIGCBench, represents a significant leap towards establishing consistent and comprehensive performance metrics [42], [46]. These frameworks facilitate a deeper understanding of how well human feedback aligns with qualitative assessments, and they help identify distinct strengths and limitations of generative models in varied contexts, fostering iterative improvements.

In terms of future directions, ongoing research must address the limitations of current methodologies, particularly in their ability to generalize across varied datasets and handle edge cases within video generation. The potential of fine-grained feedback mechanisms, as discussed in works surrounding language models and other generative tasks, holds promise for enriching the training of video models. Integrating multi-modal inputs—such as visual and textual—can further enhance the sophistication of feedback mechanisms, enabling more intricate and satisfactory user interactions.

In conclusion, the trajectory of human feedback integration in video generation is poised to redefine the creative landscape, driving developments that prioritize user engagement while simultaneously advancing algorithmic sophistication. The need for robust ethical frameworks to address privacy and bias concerns must also be underscored as practitioners and researchers forge ahead, ensuring inclusivity in the design and application of these emerging technologies. The convergence of AI advancements with practical human experience establishes a fertile ground for innovation, with significant implications for the future of automated content creation in an increasingly digital world.

References

[1] N. Aldausari, A. Sowmya, N. Marcus, and G. Mohammadi, "Video generative adversarial networks: A review," *ACM Computing Surveys (CSUR)*, vol. 55, pp. 1 – 25, 2020. 1, 3, 5, 9, 12, 14, 19

[2] N. Stiennon, L. Ouyang, J. Wu, D. M. Ziegler, R. J. Lowe, C. Voss, A. Radford, D. Amodei, and P. Christiano, "Learning to summarize from human feedback," *ArXiv*, vol. abs/2009.01325, 2020. 1, 2, 5, 9, 13, 15, 16, 19

[3] Z. Wu, Y. Hu, W. Shi, N. Dziri, A. Suhr, P. Ammanabrolu, N. A. Smith, M. Ostendorf, and H. Hajishirzi, "Fine-grained human feedback gives better rewards for language model training," *ArXiv*, vol. abs/2306.01693, 2023. 1, 5, 9, 13

[4] H. Ling and S. Fidler, "Teaching machines to describe images via natural language feedback," *ArXiv*, vol. abs/1706.00130, 2017. 2, 10, 12

[5] T. Kaufmann, P. Weng, V. Bengs, and E. Hüllermeier, "A survey of reinforcement learning from human feedback," *ArXiv*, vol. abs/2312.14925, 2023. 2, 6, 9, 10, 13, 16

[6] Z. He, X. Wang, W. Jiao, Z. Zhang, R. Wang, S. Shi, and Z. Tu, "Improving machine translation with human feedback: An exploration of quality estimation as a reward model," *ArXiv*, vol. abs/2401.12873, 2024. 2, 5, 10, 13

[7] G. I. Winata, H. Zhao, A. Das, W. Tang, D. D. Yao, S.-X. Zhang, and S. Sahu, "Preference tuning with human feedback on language, speech, and vision tasks: A survey," *ArXiv*, vol. abs/2409.11564, 2024. 2, 7, 17

[8] K. Lee, H. Liu, M. Ryu, O. Watkins, Y. Du, C. Boutilier, P. Abbeel, M. Ghavamzadeh, and S. Gu, "Aligning text-to-image models using human feedback," *ArXiv*, vol. abs/2302.12192, 2023. 3, 10, 17

[9] N. C. Chung, "Human in the loop for machine creativity," *ArXiv*, vol. abs/2110.03569, 2021. 3

[10] A. Kuznetsova, A. Talati, Y. Luo, K. Simmons, and V. Ferrari, "Efficient video annotation with visual interpolation and frame selection guidance," *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 3069–3078, 2020. 3

[11] D. Lindner and M. El-Assady, "Humans are not boltzmann distributions: Challenges and opportunities for modelling human feedback and interaction in reinforcement learning," *ArXiv*, vol. abs/2206.13316, 2022. 3, 10

[12] J. D. Chang, K. Brantley, R. Ramamurthy, D. K. Misra, and W. Sun, "Learning to generate better than your llm," *ArXiv*, vol. abs/2306.11816, 2023. 3

[13] Y. Zhang, Y. Kang, Z. Zhang, X. Ding, S. Zhao, and X. Yue, "Interactivevideo: User-centric controllable video generation with synergistic multimodal instructions," *ArXiv*, vol. abs/2402.03040, 2024. 3, 17

[14] Y.-Y. He, Z. Liu, J. Chen, Z. Tian, H. Liu, X. Chi, R. Liu, R. Yuan, Y. Xing, W. Wang, J. Dai, Y. Zhang, W. Xue, Q. fei Liu, Y.-T. Guo, and Q. Chen, "Llms meet multimodal generation and editing: A survey," *ArXiv*, vol. abs/2405.19334, 2024. 3, 17

[15] M. Kumar, M. Babaeizadeh, D. Erhan, C. Finn, S. Levine, L. Dinh, and D. Kingma, "Videoflow: A conditional flow-based model for stochastic video generation," *arXiv: Computer Vision and Pattern Recognition*, 2019. 3, 7, 10, 14

[16] S. Zhang, X. Yang, Y. Feng, C. Qin, C.-C. Chen, N. Yu, Z. Chen, H. Wang, S. Savarese, S. Ermon, C. Xiong, and R. Xu, "Hive: Harnessing human feedback for instructional visual editing," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9026–9036, 2023. 3, 14

[17] Y. Zhao, R. Joshi, T. Liu, M. Khalman, M. Saleh, and P. J. Liu, "Slic-hf: Sequence likelihood calibration with human feedback," *ArXiv*, vol. abs/2305.10425, 2023. 3, 17

[18] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, and D. Lorenz, "Stable video diffusion: Scaling latent video diffusion models to large datasets," *ArXiv*, vol. abs/2311.15127, 2023. 4

[19] B. Peng, J. Wang, Y. Zhang, W. Li, M. Yang, and J. Jia, "Controlnext: Powerful and efficient control for image and video generation," *ArXiv*, vol. abs/2408.06070, 2024. 4

[20] W.-L. Lei, J. Wang, F. Ma, G. Huang, and L. Liu, "A comprehensive survey on human video generation: Challenges, methods, and insights," *ArXiv*, vol. abs/2407.08428, 2024. 4

[21] X. He, D. Jiang, G. Zhang, M. W. Ku, A. Soni, S. Siu, H. Chen, A. Chandra, Z. Jiang, A. Arulraj, K. Wang, Q. D. Do, Y. Ni, B. Lyu, Y. Narsupalli, R. R. Fan, Z. Lyu, Y. Lin, and W. Chen, "Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation," *ArXiv*, vol. abs/2406.15252, 2024. 4, 10, 12, 15, 18

[22] J. Scheurer, J. A. Campos, T. Korbak, J. S. Chan, A. Chen, K. Cho, and E. Perez, "Training language models with language feedback at scale," *ArXiv*, vol. abs/2303.16755, 2023. 5, 15

[23] X. Wu, K. Sun, F. Zhu, R. Zhao, and H. Li, "Human preference score: Better aligning text-to-image models with human preference," *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2096–2105, 2023. 5

[24] P. Duan, J. Warner, Y. Li, and B. Hartmann, "Generating automatic feedback on ui mockups with large language models," *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2024. 5, 8

[25] L. Chen, G. Cui, Z. Kou, H. Zheng, and C. Xu, "What comprises a good talking-head video generation?: A survey and benchmark," *ArXiv*, vol. abs/2005.03201, 2020. 5

[26] H. Yuan, S. Zhang, X. Wang, Y. Wei, T. Feng, Y. Pan, Y. Zhang, Z. Liu, S. Albanie, and D. Ni, "Instructvideo: Instructing video diffusion models with human feedback," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6463–6474, 2023. 6

[27] X. Wang, H. Yuan, S. Zhang, D. Chen, J. Wang, Y. Zhang, Y. Shen, D. Zhao, and J. Zhou, "Videocomposer: Compositional video synthesis with motion controllability," *ArXiv*, vol. abs/2306.02018, 2023. 6, 11, 17, 18

[28] O. Daniels-Koch and R. Freedman, "The expertise problem: Learning from specialized feedback," *ArXiv*, vol. abs/2211.06519, 2022. 6

[29] J. Abramson, A. Ahuja, F. Carnevale, P. Georgiev, A. Goldin, A. Hung, J. Landon, J. Lhotka, T. Lillicrap, A. Muldal, G. Powell, A. Santoro, G. Scully, S. Srivastava, T. von Glehn, G. Wayne, N. Wong, C. Yan, and R. Zhu, "Improving multimodal interactive agents with reinforcement learning from human feedback," *ArXiv*, vol. abs/2211.11602, 2022. 6

[30] C. Vondrick, H. Pirsiavash, and A. Torralba, "Generating videos with scene dynamics," in *Neural Information Processing Systems*, 2016, pp. 613–621. 7

[31] D. Buschek, L. Mecke, F. Lehmann, and H. Dang, "Nine potential pitfalls when designing human-ai co-creative systems," *ArXiv*, vol. abs/2104.00358, 2021. 7, 15

[32] Y. Feng, E. Dohmatob, P. Yang, F. Charton, and J. Kempe, "Beyond model collapse: Scaling up with synthesized data requires reinforcement," *ArXiv*, vol. abs/2406.07515, 2024. 7

[33] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. E. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. J. Lowe, "Training language models to follow instructions with human feedback," *ArXiv*, vol. abs/2203.02155, 2022. 7, 8, 9, 14, 16, 18

[34] C. Ilisescu, H. A. Kanaci, M. Romagnoli, N. D. F. Campbell, and G. Brostow, "Responsive action-based video synthesis," *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2017. 7

[35] S. Sina, S. Kraus, and A. Rosenfeld, "Using the crowd to generate content for scenario-based serious-games," *ArXiv*, vol. abs/1402.5034, 2014. 7, 14

[36] P. Bontrager, W.-C. Lin, J. Togelius, and S. Risi, "Deep interactive evolution," *ArXiv*, vol. abs/1801.08230, 2018. 8

[37] D. Weissenborn, O. Täckström, and J. Uszkoreit, "Scaling autoregressive video models," *ArXiv*, vol. abs/1906.02634, 2019. 8

[38] C. Hou, G. Wei, Y. Zeng, and Z. Chen, "Training-free camera control for video generation," *ArXiv*, vol. abs/2406.10126, 2024. 8

[39] S. Ma, L. Cui, D. Dai, F. Wei, and X. Sun, "Livebot: Generating live video comments based on visual and textual contexts," in *AAAI Conference on Artificial Intelligence*, 2018, pp. 6810–6817. 8

[40] H. Lee, S. Phatale, H. Mansoor, K. Lu, T. Mesnard, C. Bishop, V. Carbune, and A. Rastogi, "Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback," in *International Conference on Machine Learning*, 2023. 8, 15

[41] T. Unterthiner, S. van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, "Towards accurate generative models of video: A new metric & challenges," *ArXiv*, vol. abs/1812.01717, 2018. 9, 10, 19

[42] F. Fan, C. Luo, W. Gao, and J. Zhan, "Aigcbench: Comprehensive evaluation of image-to-video content generated by ai," *ArXiv*, vol. abs/2401.01651, 2024. 9, 12, 19

[43] P. Fernandes, A. Madaan, E. Liu, A. Farinhas, P. H. Martins, A. Bertsch, J. G. C. de Souza, S. Zhou, T. S. Wu, G. Neubig, and A. F. T. Martins, "Bridging the gap: A survey on integrating (human) feedback for natural language generation," *ArXiv*, vol. abs/2305.00955, 2023. 9, 11, 17

[44] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. Dassarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, N. Joseph, S. Kadavath, J. Kernion, T. Conerly, S. El-Showk, N. Elhage, Z. Hatfield-Dodds, D. Hernandez, T. Hume, S. Johnston, S. Kravec, L. Lovitt, N. Nanda, C. Olsson, D. Amodei, T. B. Brown, J. Clark, S. McCandlish, C. Olah, B. Mann, and J. Kaplan, "Training a helpful and harmless assistant with reinforcement learning from human feedback," *ArXiv*, vol. abs/2204.05862, 2022. 10

[45] S. Zhou, M. L. Gordon, R. Krishna, A. Narcomey, L. Fei-Fei, and M. S. Bernstein, "Hype: A benchmark for human eye perceptual evaluation of generative models," in *Neural Information Processing Systems*, 2019, pp. 3444–3456. 10

[46] Z. Huang, Y. He, J. Yu, F. Zhang, C. Si, Y. Jiang, Y. Zhang, T. Wu, Q. Jin, N. Chanpaisit, Y. Wang, X. Chen, L. Wang, D. Lin, Y. Qiao, and Z. Liu, "Vbench: Comprehensive benchmark suite for video generative models," *2024 IEEE/CVF Conference on Computer Vision*

and Pattern Recognition (CVPR), pp. 21 807–21 818, 2023. 10, 11, 14, 15, 16, 17, 18, 19

[47] Y. Liu, X. Cun, X. Liu, X. Wang, Y. Zhang, H. Chen, Y. Liu, T. Zeng, R. Chan, and Y. Shan, "Evalcrafter: Benchmarking and evaluating large video generation models," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22 139–22 149, 2023. 11, 12, 17

[48] M. Liao, H. Lu, X. Zhang, F. Wan, T. Wang, Y. Zhao, W. Zuo, Q. Ye, and J. Wang, "Evaluation of text-to-video generation models: A dynamics perspective," *ArXiv*, vol. abs/2407.01094, 2024. 11

[49] Y. Peng, Y. Cui, H. Tang, Z. Qi, R. Dong, J. Bai, C. Han, Z. Ge, X. Zhang, and S.-T. Xia, "Dreambench++: A human-aligned benchmark for personalized image generation," *ArXiv*, vol. abs/2406.16855, 2024. 11

[50] H. Dang, L. Mecke, F. Lehmann, S. Goller, and D. Buschek, "How to prompt? opportunities and challenges of zero- and few-shot learning for human-ai interaction in creative applications of generative models," *ArXiv*, vol. abs/2209.01390, 2022. 11, 16

[51] J. Kreutzer, J. Uyheng, and S. Riezler, "Reliability and learnability of human bandit feedback for sequence-to-sequence reinforcement learning," in *Annual Meeting of the Association for Computational Linguistics*, 2018, pp. 1777–1788. 11, 15, 16

[52] B. Xiao, Q. Lu, B. Ramasubramanian, A. Clark, L. Bushnell, and R. Poovendran, "Fresh: Interactive reward shaping in high-dimensional state spaces using human feedback," *ArXiv*, vol. abs/2001.06781, 2020. 12

[53] W. Wang, H. Bai, J. tse Huang, Y. Wan, Y. Yuan, H. Qiu, N. Peng, and M. R. Lyu, "New job, new gender? measuring the social bias in image generation models," *ArXiv*, vol. abs/2401.00763, 2024. 12, 19

[54] J. Xing, M. Xia, Y. Liu, Y. Zhang, Y. Zhang, Y.-Y. He, H. Liu, H. Chen, X. Cun, X. Wang, Y. Shan, and T. Wong, "Make-your-video: Customized video generation using textual and structural guidance," *IEEE transactions on visualization and computer graphics*, vol. PP, 2023. 12

[55] Y. Oshima, S. Taniguchi, M. Suzuki, and Y. Matsuo, "Ssm meets video diffusion models: Efficient video generation with structured state spaces," *ArXiv*, vol. abs/2403.07711, 2024. 12

[56] W. Shen, X. Zhang, Y. Yao, R. Zheng, H. Guo, and Y. Liu, "Improving reinforcement learning from human feedback using contrastive rewards," *ArXiv*, vol. abs/2403.07708, 2024. 13

[57] V. Chen, A. Gupta, and K. Marino, "Ask your humans: Using human instructions to improve generalization in reinforcement learning," *ArXiv*, vol. abs/2011.00517, 2020. 13

[58] E. Heim, "Constrained generative adversarial networks for interactive image generation," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10 745–10 753, 2019. 14

[59] H. Dang, L. Mecke, and D. Buschek, "Ganslider: How users control generative models for images using multiple sliders with and without feedforward information," *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022. 14

[60] J. Xu, X. Liu, Y. Wu, Y. Tong, Q. Li, M. Ding, J. Tang, and Y. Dong, "Imagereward: Learning and evaluating human preferences for text-to-image generation," *ArXiv*, vol. abs/2304.05977, 2023. 14

[61] Y. Hu, C. Luo, and Z. Chen, "Make it move: Controllable image-to-video generation with text descriptions," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18 198–18 207, 2021. 15, 18

[62] D. Arumugam, J. K. Lee, S. Saskin, and M. Littman, "Deep reinforcement learning from policy-dependent human feedback," *ArXiv*, vol. abs/1902.04257, 2019. 15, 18

[63] X. Ma, S. Wijewickrema, S. Zhou, Y. Zhou, Z. Mhammedi, S. O'Leary, and J. Bailey, "Adversarial generation of real-time feedback with neural networks for simulation-based training," in *International Joint Conference on Artificial Intelligence*, 2017, pp. 3763–3769. 15

[64] S. Yang, L. Jiang, Z. Liu, and C. C. Loy, "Vtoonify," *ACM Transactions on Graphics (TOG)*, vol. 41, pp. 1 – 15, 2022. 16

[65] J. Li, W. Feng, T.-J. Fu, X. Wang, S. Basu, W. Chen, and W. Y. Wang, "T2v-turbo: Breaking the quality bottleneck of video consistency model with mixed reward feedback," *ArXiv*, vol. abs/2405.18750, 2024. 16

[66] D. Wang, H.-R. Wei, Z. Zhang, S. Huang, J. Xie, W. Luo, and J. Chen, "Non-parametric online learning from human feedback for neural machine translation," in *AAAI Conference on Artificial Intelligence*, 2021, pp. 11 431–11 439. 16

[67] J. Shi, R. Jain, H. Doh, R. Suzuki, and K. Ramani, "An hci-centric survey and taxonomy of human-generative-ai interactions," *ArXiv*, vol. abs/2310.07127, 2023. 17

[68] Y. Wei, S. Zhang, Z. Qing, H. Yuan, Z. Liu, Y. Liu, Y. Zhang, J. Zhou, and H. Shan, "Dream video: Composing your dream videos with customized subject and motion," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6537–6549, 2023. 17

[69] T.-S. Chen, C. Lin, H.-Y. Tseng, T.-Y. Lin, and M. Yang, "Motion-conditioned diffusion model for controllable video synthesis," *ArXiv*, vol. abs/2304.14404, 2023. 17

[70] S. Goldfarb-Tarrant, H. Feng, and N. Peng, "Plan, write, and revise: an interactive system for open-domain story generation," in *North American Chapter of the Association for Computational Linguistics*, 2019, pp. 89–97. 18

[71] S.-S. Yin, C. Wu, J. Liang, J. Shi, H. Li, G. Ming, and N. Duan, "Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory," *ArXiv*, vol. abs/2308.08089, 2023. 18

[72] Y. Wang, X. Chen, X. Ma, S. Zhou, Z. Huang, Y. Wang, C. Yang, Y. He, J. Yu, P. der Yang, Y. Guo, T. Wu, C. Si, Y. Jiang, C. Chen, C. C. Loy, B. Dai, D. Lin, Y. Qiao, and Z. Liu, "Lavie: High-quality video generation with cascaded latent diffusion models," *ArXiv*, vol. abs/2309.15103, 2023. 18

[73] Y. Zhang, Y. Wei, D. Jiang, X. Zhang, W. Zuo, and Q. Tian, "Controlvideo: Training-free controllable text-to-video generation," *ArXiv*, vol. abs/2305.13077, 2023. 18

[74] Z. Hu and D. Xu, "Videocontrolnet: A motion-guided video-to-video translation framework by using diffusion model with controlnet," *ArXiv*, vol. abs/2307.14073, 2023. 19