AllocMV: Optimal Resource Allocation for Music Video Generation via Structured Persistent State

Huimin Wang [1] **Chang Xia** [1] **Leilei Ouyang** [1] **Yongqi Kang** [1] **Yu Fu** [1] **Yuqi Ouyang** [1]

Abstract

Generating long-horizon music videos (MVs) is frequently constrained by prohibitive computational costs and difficulty maintaining cross-shot consistency. We propose AllocMV, a hierarchical framework formulating music video synthesis as a Multiple-Choice Knapsack Problem (MCKP). AllocMV represents the video's persistent state as a compact, structured object comprising character entities, scene priors, and sharing graphs, produced by a global planner prior to realization. By estimating segment saliency from multimodal cues, a group-level MCKP solver based on dynamic programming performs budget-aware allocation across High-Gen, Mid-Gen, and Reuse branches under calibrated proxy utilities. For repetitive musical motifs, we implement a divergence-based forking strategy that reuses visual prefixes to reduce costs while ensuring motif-level continuity. Evaluated via the Cost-Quality Ratio (CQR), AllocMV achieves an optimal trade-off between perceived quality and resource expenditure under strict budgetary and rhythmic constraints.

1. Introduction

Diffusion models have opened new frontiers in local video synthesis. However, extending these models for professional-level, long-term content remains a formidable challenge (Wang et al., 2023a; Qiu et al., 2024; Zhou et al., 2024). This is particularly evident in music video (MV) generation, since videos are rhythmic narratives rather than mere image sequences (Kim et al., 2022; Mao et al., 2025; Tang et al., 2025). Thus, MV generation must follow strict structural prerequisites. Unlike ordinary long videos, high-quality MVs require precise multi-modal synchronization: visual elements must match recurring musical themes, transitions must align with beats, and narrative progression must reflect the emotional prominence of lyrics within minutes (Kim et al., 2022; Mao et al., 2025; Tang et al., 2025). In professional filmmaking, these complexities are managed through strict pre-production, during which the director uses the script, character library, and scene scouting to maintain consistency within a limited budget. However, existing automated MV generation frameworks (e.g., AutoMV) often overlook this structural hierarchy, typically defaulting to a uniform resource allocation strategy (Tang et al., 2025). By treating all song segments as perceptually equivalent, these systems fail to prioritize high-impact moments like climactic choruses, redundantly generating segments that could benefit from motif-level reuse. This simplistic approach is not only inefficient in inference costs but also exacerbates the *identity drift* phenomenon, causing gradual character or scene fidelity loss, especially when the model lacks a compact, persistent representation of global creative intent (Zhou et al., 2024; Kahatapitiya et al., 2025).

Addressing these challenges, we introduce AllocMV, a framework formulating full-song MV synthesis as a Multiple-Choice Knapsack Problem (MCKP), with the pipeline illustrated in Figure 1. For structured content like MVs, we propose the persistent state be an explicit, executable object produced by a global planner before realization. Via a two-round dynamic programming algorithm, AllocMV optimally assigns segments to *High-Gen*, *Mid-Gen*, or *Reuse* branches, concentrating resources where most perceptually salient. Our contributions are summarized below:

- **Structured Persistent State:** We formalize video persistent state as a compact, executable object that bundles character identities, scene priors, and sharing graphs, effectively decoupling long-range consistency from the stochastic nature of diffusion models.
- **Hierarchical Budget-Aware Planning:** We model MV generation as an MCKP and introduce a two-round dynamic programming algorithm for global optimal resource allocation under non-local dependency constraints.
- **Divergence-based Forking Strategy:** We implement a novel motif-reuse mechanism that utilizes shared visual prefixes and divergent suffix generation to achieve

[1]College of Computer Science, Sichuan University. Correspondence to: Yuqi Ouyang <yuqi.ouyang@scu.edu.cn>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

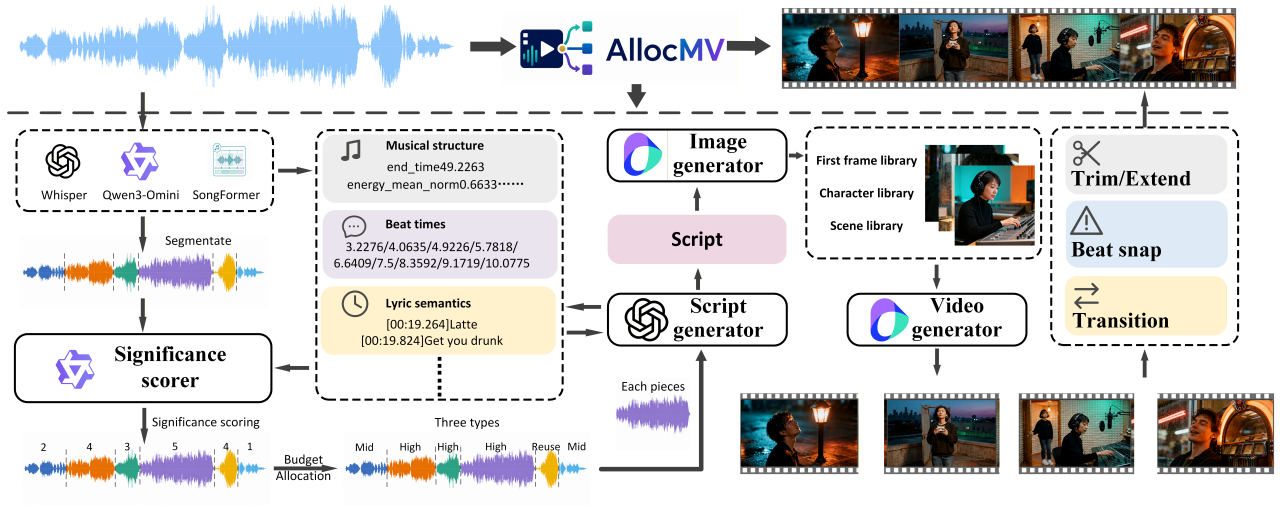

Figure 1. Overview of AllocMV. Given an input song, the system extracts musical structure, beats, lyrics, and saliency cues, performs global script planning, and routes each segment to a High-Gen, Mid-Gen, or Reuse branch under a fixed budget. Generated segments are finally combined with beat-synchronized assembly to produce the full MV.

high-fidelity motif consistency while minimizing redundant computation.

- **Cost-Aware Evaluation for Long-Horizon Video:** We introduce the Cost-Quality Ratio (CQR), a unified quality-to-cost metric for evaluating the efficiency of structured long-horizon video generation under computational budget constraints.

2. Method

2.1. Problem Formulation

We define the long-horizon MV generation task as a structured narrative optimization problem. Given an input tuple $X = \{A, T, M\}$, where A denotes the acoustic stream, T the lyric text, and M the associated metadata, we first decompose X into a sequence of N contiguous segments $X = \{x_1, \ldots, x_N\}$. For each segment x_i, we associate two attributes: a duration $d_i \in \mathbb{R}^+$ measuring its temporal length within the song, and a perceptual saliency weight $m_i \in \mathbb{R}^+$ scoring its narrative importance from multimodal cues (lyric semantics and acoustic energy). We then introduce a discrete action set $O = \{\textit{High}, \textit{Mid}, \textit{Reuse}\}$ and formulate a planning problem that assigns an action $o_i \in O$ to each segment under the global budget constraint B. The optimal action sequence $\mathbf{o} = \{o_1, \ldots, o_N\}$ is obtained by solving:

$$\max_{o_i \in O} \sum_{i=1}^{N} m_i \cdot d_i \cdot Q(o_i) \quad \text{s.t.} \quad \sum_{i=1}^{N} C(o_i, d_i) \leq B, \quad (1)$$

where $Q(o_i)$ denotes the quality of action o_i, and $C(o_i, d_i)$ is its computational cost. Given the optimized plan, the final MV is generated in an execution stage conditioned on persistent states s_i which encodes shared entities and cross-segment structural priors, denoted as $v_i = \mathcal{G}(s_i, o_i)$, producing the final video sequence $V = \{v_1, \ldots, v_N\}$.

2.2. Persistent Narrative State

To mitigate identity drift in long-horizon generation, we introduce an explicit persistent state $S = \{I, E, G, M, O\}$, which decouples long-range consistency from the stochastic nature of the generative model. Here, I and E denote identity and environment libraries that encode global character identities and environmental priors, serving as consistency anchors across segments. The narrative sharing graph G defines a directed topology for the propagation of visual motifs, modeling owner–consumer relationships between segments to ensure motif-level continuity. The motif index set M and action assignments O maintain compact references to reusable assets and optimized production choices, respectively. A key property of S is its auditability, allowing the full narrative structure to be inspected or edited prior to rendering, thereby enabling controllable and interpretable long-horizon generation.

2.3. System Architecture Overview

AllocMV employs a decoupled modular pipeline designed to ensure that system performance scales with the capabilities of the underlying foundation models. As depicted in Figure 1, the architecture consists of five sequential phases: (i) *Multimodal Structural Analysis*, where Whisper, Qwen3-Omni, and SongFormer jointly extract word-level lyric timestamps, musical structure with normalized energy curves, and beat anchors, partitioning the song into N contiguous segments; These analyzers are used

as replaceable front-end modules: AllocMV only requires segment-level saliency, beat anchors, and optional structural labels, rather than depending on a specific music-analysis model. (ii) *Strategic Planning*, where an LLM-based significance scorer assigns each segment a perceptual saliency m_i from multimodal cues, after which a group-level multiple-choice knapsack solver maps every segment to one of $\{High, Mid, Reuse\}$ under the global budget B, and a script generator produces shot-level prompts for the resulting plan; (iii) *Visual Asset Initialization*, where an image generator, conditioned on the script, instantiates the persistent state S as a character library, a scene library, and a first-frame library that serve as cross-segment identity anchors; (iv) *Hierarchical Video Synthesis*, where a video generator renders per-segment clips from the prompts and library assets, with sampling cost and fidelity determined by the tier a_i assigned in phase (ii); (v) *Temporal Narrative Assembly*, where Trim/Extend, beat snapping to acoustic accents, and prior-driven transition synthesis stitch the clips into a beat-synchronized, coherent long-form MV.

2.4. Global Resource Planning via Group-Level MCKP

The sharing graph introduces non-local dependencies between segments, making per-segment decisions suboptimal. We formulate this as a group-level MCKP(Sinha & Zoltners, 1979). Let G be the set of sharing groups. For each group $g \in G$, there exists a set of candidate plans P where each plan $p \in P$ specifies a joint action configuration for all members within g. We then define binary decision variables $y_{gp} \in 0, 1$, where $y_{gp} = 1$ indicates selecting plan p for group g. The global optimization problem is formulated as:

$$\begin{aligned} \max_{y_{gp}} \quad & \sum_{g\in G}\sum_{p\in P} U_{gp}\, y_{gp} \\ \text{s.t.} \quad & \sum_{g\in G}\sum_{p\in P} c_{gp}\, y_{gp} \le B, \\ & \sum_{p\in P} y_{gp} = 1, \quad \forall g \in G \end{aligned} \tag{2}$$

where U_{gp} denotes the saliency-weighted utility of plan p, c_{gp} is the corresponding computational cost. The first constraint enforces the global budget limit, while the second constraint enforces selection uniqueness, requiring exactly one plan to be chosen for each group. This integer program is solved via dynamic programming to achieve a globally optimal trade-off between narrative consistency and budget efficiency.

2.5. Temporal Anchor Control and Narrative Smoothing

During *Phase (v): Temporal Narrative Assembly* (see Section 2), musical beats serve as discrete temporal anchors for the visual narrative, achieving audiovisual synchrony. Segment boundaries are aligned with the acoustic accents and downbeats extracted during *Phase (i): Multimodal Structural Analysis*, reinforcing rhythmic coherence throughout the generated MV.

3. Experiments

Data. We curate a pilot benchmark consisting of five full-length songs spanning five genres: pop, rock, ballad, electronic, and folk. Each song contains clearly identifiable structural segments, including intro, verse, chorus, bridge, and outro sections. The mean track duration is 94 ± 11 s.

Implementation Details. All methods in this work share the same preprocessing and generation pipeline for fair comparison. Word-level lyric timestamps are obtained using Whisper (Radford et al., 2022), while structural segmentation and downbeat are derived from SongFormer (Hao et al., 2025), fused with the Qwen3-Omni (Xu et al., 2025) energy curve. Per-segment saliency weights $m_i \in \{1, \ldots, 5\}$ are estimated using Qwen-Plus (Yang et al., 2025). For video synthesis, we employ Seedance (Gao et al., 2025b) as the generation backend, with Seedream-generated identity and environment priors as anchors for the persistent state S (Gao et al., 2025a). Computational cost is reported in USD, derived from both backend video generation expenses and auxiliary large model calls. The average full *High-Gen* configuration yields a maximum cost of $B_{\max} = 2.85$ USD per song. We therefore set the operating budget to $0.6B_{\max} \approx 1.71$ USD per song for all budget-constrained methods, while *Uniform-High* is included as an unconstrained upper-bound reference. Following (Zheng et al., 2023), we parameterize the MCKP quality factors $Q(\cdot)$ using a VLM-as-a-Judge protocol. A frozen GPT-4o model evaluates generated samples on a five-point scale across four dimensions: semantic alignment, audio-visual consistency, rhythmic precision, and motif stability. The final proxy quality values $Q(High)$, $q(Mid)$, and $Q(Reuse)$ are computed as the mean score over these dimensions. Note that these calibrated values are used exclusively for optimization within the MCKP framework and are distinct from the post-hoc evaluation metrics used for final performance reporting, such as ImageBind and CLIP-Score.

Evaluation Metrics. We evaluate generation quality using ImageBind Score (Girdhar et al., 2023) for audio-visual semantic alignment, CLIP-Score (Hessel et al., 2021) for text-video consistency, BeatAlign (Li et al., 2021) for rhythmic synchronization, and Motif Consistency to assess identity stability, measured via SSIM or LPIPS across recurring visual motifs (Wang et al., 2004; Zhang et al., 2018). For efficiency evaluation, we report total Cost in USD and the proposed **Cost-Quality Ratio (CQR)**, defined as saliency-weighted per-segment quality normalized by total genera-

tion cost:

$$\text{CQR} = \frac{\sum_{i=1}^{N} m_i \cdot Q_i}{\sum_{i=1}^{N} C(o_i^*)}, \tag{3}$$

where $m_i \in [0, 1]$ is the normalized perceptual saliency of segment i (Section 2.2), $C(o_i^*)$ is its amortized generation cost in USD, and $Q_i \in [1, 5]$ is the calibrated segment-level quality score produced by the VLM judge. Following the branch-utility calibration used by the MCKP solver, raw per-segment quality estimates are grouped by assigned action and normalized. Higher CQR indicates better quality per unit cost under the optimized proxy utility.

3.1. Comparisons with Previous Methods

We compare AllocMV against MuseV (Wang et al., 2023b), VideoComposer (Wang et al., 2023c), and AutoMV (Tang et al., 2025) on the same five-song benchmark under identical SongFormer-derived segmentation for fair evaluation. As shown in Table 1, prior methods achieve competitive per-frame quality but fail to capture long-horizon structure, leading to consistently weak BeatAlign around 0.18 and higher overall cost. In contrast, AllocMV delivers significantly stronger temporal and structural coherence, achieving the highest BeatAlign 0.6679 while maintaining competitive CLIP 0.3014 and the best overall CQR 0.7586 under the same budget constraint. Compared with AutoMV, which obtains slightly higher CLIP 0.3222, AllocMV improves rhythmic alignment by more than 0.55 in absolute BeatAlign and motif consistency by 0.146, while reducing cost by 48 percent. Overall, AllocMV consistently achieves the best trade-off between quality, structure, and efficiency across all evaluated systems.

Method	BeatAlign ↑	CQR ↑	CLIP ↑	Motif ↑	Cost ↓
MuseV	$0.0831_{\pm.021}$	$0.2083_{\pm.028}$	$0.2512_{\pm.019}$	$0.8812_{\pm.024}$	$3.04_{\pm.19}$
VideoComposer	$0.1024_{\pm.024}$	$0.2210_{\pm.031}$	$0.2318_{\pm.022}$	$0.8754_{\pm.026}$	$3.15_{\pm.21}$
AutoMV	$0.0960_{\pm.023}$	$0.4697_{\pm.036}$	$\mathbf{0.3222}_{\pm.017}$	$0.8521_{\pm.029}$	$3.25_{\pm.22}$
AllocMV (Ours)	$\mathbf{0.6679}_{\pm.039}$	$\mathbf{0.7586}_{\pm.034}$	$0.3014_{\pm.018}$	$\mathbf{0.9984}_{\pm.0008}$	$\mathbf{1.69}_{\pm.10}$

Table 1. Comparison with state-of-the-art MV generation baselines. Mean ± std over N=5 songs.

3.2. Ablation Studies

We evaluate AllocMV against three baseline settings and three ablated variants, as summarized in Table 2. The baselines include *Uniform-Mid* and *Uniform-High*, which generate all segments using a single fixed tier without persistent state modeling or beat synchronization, and *Heuristic*, which assigns High-Gen to chorus and bridge sections while using Mid-Gen elsewhere with a simple duplicate-suppression rule. The three ablations further isolate key components: *w/o Budget Allocation* replaces the group-level MCKP with heuristic assignment, *w/o Beat-Sync Assembly* removes beat-aligned temporal snapping, and *w/o Motif Reuse* disables the sharing graph G, enforcing independent generation across segments.

The results highlight several key effects. Among budget-compliant methods, AllocMV achieves the highest CQR (0.7586), improving over the strongest *Heuristic* baseline (0.6943) by 9.3% while using 12% less cost. Uniform strategies further illustrate the trade-off between allocation and quality: *Uniform-High* violates the budget by 66% and drops to CQR 0.5754, while *Uniform-Mid* remains budget-efficient yet is limited to 0.7406 due to uniform under-allocation. In the ablations, removing budget allocation reduces CQR from 0.7586 to 0.7034, indicating a clear loss in cost-aware efficiency despite relatively stable CLIP performance. Disabling beat-synchronized assembly causes a dramatic drop in BeatAlign from 0.6679 to 0.1830, showing that rhythmic alignment collapses without temporal anchoring. Finally, removing motif reuse increases cost from 1.69 to 2.10 while degrading motif consistency from 0.9984 to 0.9210, demonstrating reduced efficiency and weakened identity preservation. These results confirm that the proposed components contribute in a complementary manner, jointly enabling efficient and structurally coherent long-horizon generation.

Method	CQR ↑	B-Align ↑	CLIP ↑	Motif ↑	Cost ↓
Uniform-Mid	$0.7406_{\pm0.038}$	$0.1820_{\pm0.045}$	$0.2451_{\pm0.018}$	$0.8512_{\pm0.028}$	$1.55_{\pm0.12}$
Uniform-High	$0.5754_{\pm0.029}$	$0.1820_{\pm0.045}$	$0.2754_{\pm0.020}$	$0.8623_{\pm0.026}$	$2.85_{\pm0.18}$
Heuristic	$0.6943_{\pm0.041}$	$0.1815_{\pm0.044}$	$0.2820_{\pm0.019}$	$0.8721_{\pm0.025}$	$1.92_{\pm0.15}$
w/o Allocation	$0.7034_{\pm0.039}$	$0.6650_{\pm0.039}$	$0.2952_{\pm0.018}$	$0.9971_{\pm0.0012}$	$1.85_{\pm0.14}$
w/o Beat-Sync	$\mathbf{0.7586}_{\pm0.034}$	$0.1830_{\pm0.047}$	$0.3010_{\pm0.018}$	$\mathbf{0.9984}_{\pm0.0008}$	$\mathbf{1.69}_{\pm0.10}$
w/o Reuse	$0.6480_{\pm0.043}$	$\mathbf{0.6710}_{\pm0.040}$	$\mathbf{0.3015}_{\pm0.020}$	$0.9210_{\pm0.022}$	$2.10_{\pm0.15}$
AllocMV (Ours)	$\mathbf{0.7586}_{\pm0.034}$	$\mathbf{0.6679}_{\pm0.039}$	$\mathbf{0.3014}_{\pm0.018}$	$\mathbf{0.9984}_{\pm0.0008}$	$\mathbf{1.69}_{\pm0.10}$

Table 2. Quantitative comparison against baselines and ablations under fixed budget B. Mean ± std over N=5 songs.

4. Discussion and Conclusion

This work identifies a key limitation in long-horizon video generation: the absence of an explicit, executable state representation at the system level beyond the capacity of foundation models. We propose AllocMV, which introduces a structured persistent state that compresses cross-segment information into a planner-driven object, thereby decoupling global consistency from the stochastic nature of diffusion generation. The results suggest that effective long-horizon generation requires explicit state design to maintain narrative coherence under budget constraints, hierarchical control to bridge global planning with segment-level generation decisions, and system-level evaluation that captures both efficiency and structural stability, as reflected in metrics such as motif consistency and the designed CQR and Motif Consistency. While AllocMV is effective for structured music video generation, our current evaluation remains a pilot study over five full-length songs and should be further validated on larger and more diverse music collections. In ad-

dition, the pipeline benefits from explicit musical structure, lyric timestamps, and beat anchors extracted by external models, and may be less reliable for free-form audio or narratives without clear rhythmic structure. Finally, the MCKP objective relies on calibrated proxy utilities, including VLM-based quality estimates, which may not fully capture human aesthetic preferences. Future work will extend the benchmark scale, incorporate more robust audio-structure analysis, and combine proxy-based optimization with broader human preference evaluation.

References

Gao, Y., Gong, L., Guo, Q., Hou, X., Lai, Z., Li, F., Li, L., Lian, X., et al. Seedream 3.0 technical report, 2025a. URL https://arxiv.org/abs/2504.11346.

Gao, Y., Guo, H., Hoang, T., Huang, W., Jiang, L., Kong, F., Li, H., Li, J., et al. Seedance 1.0: Exploring the boundaries of video generation models, 2025b. URL https://arxiv.org/abs/2506.09113.

Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K. V., Joulin, A., and Misra, I. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15180–15190, 2023.

Hao, C., Yuan, R., Yao, J., Deng, Q., Bai, X., Xue, W., and Xie, L. Songformer: Scaling music structure analysis with heterogeneous supervision, 2025. URL https://arxiv.org/abs/2510.02797.

Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., and Choi, Y. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528, 2021.

Kahatapitiya, K., Liu, H., He, S., Liu, D., Jia, M., Zhang, C., Ryoo, M. S., and Xie, T. Adaptive caching for faster video generation with diffusion transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://arxiv.org/abs/2411.02397.

Kim, Y., Jang, J., and Shin, S. Music2video: Automatic generation of music video with fusion of audio and text, 2022. URL https://arxiv.org/abs/2201.03809.

Li, R., Yang, S., Ross, D. A., and Kanazawa, A. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13401–13412, 2021.

Mao, Z., Zhao, M., Wu, Q., Zhong, Z., Liao, W.-H., Wakaki, H., and Mitsufuji, Y. Cross-modal learning for music-to-music-video description generation, 2025. URL https://arxiv.org/abs/2503.11190.

Qiu, H., Xia, M., Zhang, Y., He, Y., Wang, X., Shan, Y., and Liu, Z. Freenoise: Tuning-free longer video diffusion via noise rescheduling. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ijoqFqSC7p.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. Robust speech recognition via large-scale weak supervision, 2022. URL https://arxiv.org/abs/2212.04356.

Sinha, P. and Zoltners, A. A. The multiple-choice knapsack problem. *Operations Research*, 27(3):503–515, 1979. doi: 10.1287/opre.27.3.503.

Tang, X., Lei, X., Zhu, C., Chen, S., Yuan, R., Li, Y., Oh, C., Zhang, G., Huang, W., Benetos, E., Liu, Y., Liu, J., and Ma, Y. Automv: An automatic multi-agent system for music video generation, 2025. URL https://arxiv.org/abs/2512.12196.

Wang, F.-Y., Chen, W., Song, G., Ye, H.-J., Liu, Y., and Li, H. Gen-l-video: Multi-text to long video generation via temporal co-denoising, 2023a. URL https://arxiv.org/abs/2305.18264.

Wang, X., Shi, Y., et al. Musev: Infinite-length and high fidelity virtual human video generation with visual conditioned parallel denoising. *arXiv preprint arXiv:2309.08461*, 2023b.

Wang, X., Zhang, H., et al. Videocomposer: Compositional video synthesis with motion controllability. In *Advances in Neural Information Processing Systems*, volume 36, 2023c.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

Xu, J., Guo, Z., Hu, H., Chu, Y., Wang, X., He, J., Wang, Y., Shi, X., et al. Qwen3-omni technical report, 2025. URL https://arxiv.org/abs/2509.17765.

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., et al. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, H., Lin, Z., Li, Z., Li, D., Xing, E., Zhang, H., Gonzalez, J. E., and Stoica, I. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623, 2023.

Zhou, Y., Zhou, D., Cheng, M.-M., Feng, J., and Hou, Q. Storydiffusion: Consistent self-attention for long-range image and video generation. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3501. URL https://papers.nips.cc/paper_files/paper/2024/hash/c7138635035501eb71b0adf6ddc319d6-Abstract-Conference.html.