ChatGPT

Pesquisa de Estado da Arte em Geração de Vídeo Inteligente

Para situar o ADUC-SDR, examinamos os trabalhos recentes que visam gerar vídeos longos e coerentes. Huang et al. (WACV 2025) propõem um pipeline multi-etapa: geram **quadros-chave** representativos de eventos-chave e depois interpolam frames intermediários [1]. Esse modelo simplifica vídeos longos prevendo keyframes que condensam conceitos de alto nível [1], usando sinais como rótulos de objeto e layouts visuais como orientação [2]. Embora supere métodos convencionais, **não usa nem LLM nem "estado" dinâmico entre segmentos**: ela assume que todos os quadros necessários (passado e futuro imediato) são fornecidos a cada etapa. Os autores mesmos ressaltam que vídeos multi-trecho produzidos (ex. VLOGGER, LVD) "carecem de continuidade temporal direta entre cenas" e que gerar **vídeo longo em um único "take"** com eventos variados permanece desafiador [3]. Em suma, abordagens por keyframes "empacotam" eventos importantes, mas tratam cada segmento isoladamente sem memória externa.

Diversos trabalhos focam também em **interpolação entre keyframes**. Por exemplo, métodos de "motion in-betweening" sintetizam transições entre quadros-chave passados e futuros, frequentemente baseados em difusão ou splines [4]. No entanto, esses modelos são voltados a movimentos corporais (animação de personagens) ou trechos curtos, sem um módulo de alto nível que planeja história ou reutiliza estados entre segmentos.

LLMs como Diretores de Vídeo

Uma linha paralela usa LLMs para orquestrar geração visual. Song et al. (2024) apresentam o **DirectorLLM** [5] [6]: um LLM especializado (Llama 3) é treinado/fine-tunado para gerar sequências de pose humanas discretas a partir de prompts textuais [5] [6]. Esse "diretor" gera poses (1 FPS) que são interpoladas por um difusor linear (30 FPS) e então injetadas num gerador de vídeo (VideoCrafter+ControlNet). Ou seja, **o LLM cuida da compreensão da cena e do planejamento de movimento**, transferindo sinal de pose ao gerador. Apesar de alcançar vídeos humanos realistas, **também não há uso de "estado cinético" extraído de trechos anteriores nem de condicionamento em keyframes futuros**. Esse modelo ilustra apenas que LLMs podem produzir layouts ou poses (como também fazem métodos como Free-Bloom ou VideoDirectorGPT [6]), mas sempre para um único fluxo de vídeo, sem arquiteturas recursivas de estado.

De modo semelhante, patentes recentes (por exemplo, Song et al. 2024) descrevem sistemas que treinam LLMs para emitir tokens vetorizados (via VQ-VAE) que representam pose humana [6]. Essas arquiteturas *decouplam* tarefas – o LLM planeja dinamicamente, a rede de difusão interpola e o gerador final renderiza. **Não encontramos, porém, nada que transporte explicitamente o "resíduo" (cálculo de estado) de um trecho de vídeo para o próximo**, nem que use um LLM para gerenciar múltiplos modelos via um loop de feedback complexo.

Memória de Contexto em Vídeos Longos

Outra abordagem investiga **uso de memória explícita** em geração de vídeo sequencial. Yu et al. (2025) propõem o método *Context-as-Memory* [7] para vídeos longos interativos (ex.: game streaming). Eles armazenam quadros históricos "como memória" (na forma de imagens) e condicionam cada novo frame incluindo esses quadros anteriores diretamente na entrada do modelo [7]. Além disso, usam um módulo de Recuperação de Memória que seleciona apenas frames relevantes (via sobreposição de campo de visão) [7]. Esse sistema alcança maior consistência de cena em longos vídeos, mas **depende puramente de concatenar quadros anteriores**, sem LLM ou meta-agente orquestrando nada. Ou seja, embora use "contexto" extenso, difere do eco dinâmico do ADUC: não extrai estatísticas ou vetores de estado reduzidos (só insere imagens passadas), nem aproveita preview de keyframes futuros. Outros trabalhos em vídeo iterativo (ex. em ambientes 3D) seguem linha parecida, usando janelas deslizantes de frames ou fusões de últimos frames, mas tipicamente limitados a algumas dezenas de frames.

Ferramentas Comerciais Recentes e Modelos Avançados

Algumas soluções de ponta focam em **continuidade visual**, mas sem nossos mecanismos específicos. Runway Gen-4 (2025) — modelo de difusão de vídeo multimodal — permite ao usuário fornecer uma imagem de referência e gera vídeos em múltiplos takes mantendo consistência de personagem e cena [8]. Lightricks lançou os modelos LTXV (2025) com geração rápido e alta consistência visual; o modelo LTXV-2B, por exemplo, explicitamente **suporta workflows multi-keyframe** [9]. Ambos mostram que hoje é possível produzir vídeos coesos em várias tomadas; no entanto, essas ferramentas **não descrevem nem implementam o uso de um "echo" entre segmentos nem de um LLM orquestrador**: elas funcionam internamente como pipelines de difusão multi-etapa otimizadas, sem transferir estado de um clip ao outro.

Por fim, destaca-se que os principais componentes do ADUC-SDR só surgiram muito recentemente. Por exemplo, o Gemini 1.5 Pro (fev.2024) foi o primeiro LLM multimodal a oferecer *janela de contexto* de até 1 milhão de tokens [10] (≈1 hora de vídeo), permitindo de fato raciocinar sobre longas sequências. Modelos modernos de composição de imagem (tipo "Flux.1-Kontext") e de vídeo em tempo real (LTX-Video) foram liberados apenas em 2024/2025. Antes disso, era **computacionalmente impraticável** para pesquisadores independentes orquestrar LLM + geração de imagem + vídeo em escala. Em síntese, embora haja vários grupos atacando o problema de continuidade em vídeo gerado (inclusive menções na imprensa e patentes de startups/labs), não identificamos nenhuma publicação ou implementação que combine * exatamente * nossos elementos centrais: (a) vetores de estado herdados do fim de cada trecho ("eco cinético"), (b) condicionamento por keyframes futuros ("déjà vu"), e (c) orquestração via LLM de grande contexto. Os trabalhos citados capturam aspectos isolados (keyframes, memória por concatenação, ou LLMs para poses), mas a arquitetura ADUC-SDR que governa múltiplos modelos cooperativamente permanece inédita.

Referências: Huang et al. (WACV'25) [1] [3]; Song et al. (DirectorLLM'24) [5] [6]; Yu et al. (Context-as-Memory'25) [7]; *The Verge*, Runway Gen-4 (2025) [8]; Lightricks LTXV (2025) [9]; Google Gemini 1.5 (2024) [10].

---

1 2 3 Generating Long-Take Videos via Effective Keyframes and Guidance
https://openaccess.thecvf.com/content/WACV2025/papers/Huang_Generating_Long-Take_Videos_via_Effective_Keyframes_and_Guidance_WACV_2025_paper.pdf

4 Scalable Motion In-betweening via Diffusion and Physics-Based Character Adaptation
https://arxiv.org/html/2504.09413v1

5 6 DirectorLLM for Human-Centric Video Generation
https://arxiv.org/html/2412.14484v1

7 Context as Memory: Scene-Consistent Interactive Long Video Generation with Memory Retrieval
https://arxiv.org/html/2506.03141v1

8 Runway says its latest AI video model can actually generate consistent scenes and people | The Verge
https://www.theverge.com/news/640821/runway-gen-4-artificial-intelligence-video-generator-filmmaking

9 Introducing LTXV: Our Fastest Open-Source Video Models Yet | LTX Studio
https://ltx.studio/blog/ltxv-models

10 Introducing Gemini 1.5, Google's next-generation AI model
https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/