大規模動画生成モデルにおける「時間的慣性」の適応的制御：Adaptive CFG と Temporal Unlearning による動的状態遷移の実現

著者名
所属

2026 年 2 月 14 日

概 要

HunyuanVideo や FramePack などの近年の大規模動画生成モデルは、極めて高い時間的一貫性（Temporal Consistency）を実現している。しかし、この特性は「物体が消失する」あるいは「静止から急激に跳躍する（バク転）」といった非連続的な状態遷移において、初期状態を過剰に維持しようとする **"Static Death"（静的死）** と呼ばれる現象を引き起こす。本研究では、この課題の本質がモデルの持つ過剰な「時間的慣性」にあると定義し、推論時の動的介入によってこれを制御する手法を提案する。具体的には、(1) 生成初期に強力な推力を与える **Adaptive CFG** (β)、(2) その介入を適切に減衰させ構造崩壊を防ぐ **Decay Power** (p)、(3) 過去のフレームへの固執を緩和する **Temporal Blur** の 3 要素を導入する。実験の結果、これらのパラメータ制御により、従来モデルでは不可能であった動的アクションの生成を、映像の破綻を防ぎつつ実現できることを定量的に実証した。

1 はじめに

動画生成技術の進展において、最大のマイルストーンは「フレーム間の一貫性」の獲得であった。HunyuanVideo [?] 等の最新モデルは、前のフレームの情報を強く参照することで、フリッカー（ちらつき）のない滑らかな映像生成を可能にしている。しかし、我々の予備実験において、この強力な一貫性維持機構が、特定のタスクにおいて致命的な障害となることが判明した。例えば「バク転（Backflip）」のような、静止状態から急激に身体を回転させる動作を指示した場合、モデルはプロンプトに従おうとしつつも、直前の「立っている状態」を維持しようとする慣性（Inertia）に負け、結果として微動だにしない映像が生成される。我々はこの現象をモデルの **"Static Death"** と呼ぶ。

本研究の目的は、モデルの再学習を行うことなく、推論時のパラメータ制御のみによってこの「時間的慣性」を動的に調整し、一貫性と可塑性（Plasticity）のトレードオフを解消することである。

2 関連研究

2.1 動画生成における一貫性の向上

ControlNet や AnimateDiff など、既存の多くの研究は「一貫性の欠如（フリッカー）」を解消することに主眼を置いてきた。これらは主に動きを滑らかにする、あるいは抑制する方向のアプローチであり、「動き出せないものを動かす」という課題には直接対処していない。

2.2 推論時の動的介入

Dynamic CFG や FreeInit などは、推論ステップごとにパラメータを変化させることで画質や一貫性を向上させている。しかし、これらもまた「静的な安定性」を志向しており、本研究が目指す「動的遷移（Action Initiation）」のトリガーとしては機能不全であった。

2.3 拡散モデルの周波数特性

拡散モデルの生成過程は、初期段階で低周波成分（大まかな構図・動き）が決定され、終盤で高周波成分（細部・質感）が決定されることが知られている [?]。FreeInit 等の手法は、初期ステップの潜在変数を固定・反復することで、この低周波成分の一貫性を強化しようとした。対して本研究のアプローチは、この特性を逆手に取るものである。「バク転」のようなグローバルな構造変化を伴うアクションにおいては、まさにこの初期段階（低周波領域）においてこそ、既存の「自立する人物」という構造バイアスを破壊し、新たな構造への遷移を強制する「インパルス」が必要であるという立場をとる。提案手法における β と p は、この低周波領域への介入強度と期間を明示的に制御するパラメータである。

3 提案手法

本研究では、”Static Death” を克服するために、拡散モデルの推論プロセスに介入する 2 つの手法、**Adaptive CFG (Relaxation)** と **Temporal Unlearning** を提案する。

3.1 Adaptive CFG による初期緩和

通常の Classifier-Free Guidance (CFG) におけるノイズ予測 $\hat{\epsilon}_t$ は、条件付き予測 $\epsilon_\theta(x_t, c)$ と無条件予測 $\epsilon_\theta(x_t, \emptyset)$ を用いて以下の式で表される：

$$\hat{\epsilon}_t = \epsilon_\theta(x_t, \emptyset) + s \cdot (\epsilon_\theta(x_t, c) - \epsilon_\theta(x_t, \emptyset)) \quad (1)$$

ここで s は定数のガイダンススケールである。本手法では、初期段階の過剰な拘束（静的バイアス）を緩和するため、この s をノイズレベル $\sigma_t \in [0, 1]$（1 は初期、0 は終端）の関数 $s(\sigma_t)$ として再定義する。提案する **Relaxation Curve** は以下の通りである：

$$s(\sigma_t) = s_{min} + (s_{base} - s_{min}) \cdot (1 - \beta \cdot \sigma_t^p) \quad (2)$$

各パラメータの役割は以下の通りである。

- s_{base}: ターゲットとする基本 CFG スケール（例: 6.0）。
- s_{min}: 緩和時の最小 CFG スケール（例: 1.0）。
- $\beta \in [0, 1]$ (**Relaxation Strength**): 初期ガイダンスの減衰率。$\beta = 1.0$ のとき、初期スケールは s_{min} まで低下し、モデルの探索自由度を最大化する。
- p (**Decay Power**): 通常スケールへの復帰速度を制御する指数。$p > 1$（例: 2.0）とすることで、初期の緩和状態から急速に s_{base} へ復帰させ、後半の構造生成における整合性を担保する。

3.2 Temporal Unlearning

Prompt-to-Prompt [?] 等でも示されているように、生成モデルにおける物体の同一性は Self-Attention マップに強く依存する。したがって、動画生成において直前のフレームへの Attention が強すぎることは、新しい動作への遷移を阻害する要因となる。この過剰な参照（Ghosting）を抑制するため、Self-Attention 層の Key (K) および Value (V) に対して時間軸方向の平滑化を適用する。時刻 t におけるフレーム特徴量を z_t としたとき、Key K は通常 $K = W_k z$ で計算されるが、本手法ではこれにガウシアンフィルタ $G(\cdot; \sigma_{blur})$ を適用する：

$$\tilde{K}_\tau = \sum_j G(j - \tau; \sigma_{blur}) \cdot K_j \quad (3)$$

同様に $\tilde{V}$ も計算する。これにより、Attention における時間的な局所性を強制的に下げ、直前のフレー

ム（立ち姿）への過度な Attention 集中を物理的に拡散させる。この操作により、プロンプトによる新しい状態への遷移（Overwrite）が容易になる。

4 実験・考察

4.1 検証の目的と初期仮説の修正

本実験の当初の目的は、初期段階で強力なガイダンスを与える（Boost, $\beta < 0$）ことで静的慣性を打破できるという仮説の検証であった。しかし、予備実験の結果、強力なブーストは映像の構造的崩壊を招くことが判明した（後述の Forced 条件）。これを受け、我々は逆に「初期段階のガイダンスを弱める（Relaxation, $\beta > 0$）」ことで、モデルの拘束を緩め、局所最適解（直立状態）からの脱出を促すアプローチ有効である可能性を検証した。

4.2 実験設定

HunyuanVideo を用い、難易度の高い「Backflip（バク転）」タスクにおいて評価を行った。

- **Text-Video Alignment (CLIP Score)**: プロンプトと生成動画の意味的な類似度。
- **Visual Consistency (LPIPS)**: 映像の安定性。0.5 以上は崩壊とみなす。
- **Action Recognition Impact (VideoMAE Score)**: 動作認識モデルによるターゲット動作（somersaulting, gymnastics 等）の予測確率。

4.3 定量的評価結果

以下の 3 条件での比較結果を表**??**に示す。

1. **Baseline (Static)**: 標準設定 ($\beta = 0$).
2. **Forced Boost (Collapse)**: 初期推力強化 ($\beta = -1.0$).
3. **Ours (Relaxation)**: 最適緩和設定 ($\beta = 0.75, p = 0.7$).

表 1: バク転タスクにおける定量的比較。Ours ($\beta = 0.75$）において、VideoMAE スコアが大幅に向上し、Top Class として ”somersaulting”（宙返り）が正しく検出された。

Condition	CLIP ↑	LPIPS ↓	VideoMAE ↑
Baseline ($\beta = 0$)	0.152	**0.166**	0.004
Forced ($\beta = -1.0$)	0.154	0.518	0.007
Ours ($\beta = 0.75$)	**0.187**	0.263	**0.394**

4.4 考察と結論

4.4.1 Static Death の完全な打破

表**??**が示す結果は決定的である。初期に強制的な推力を与える Forced 条件（$\beta = -1.0$）が映像の崩壊を招いたのに対し、初期ガイダンスを緩和する Ours ($\beta = 0.75$）は、VideoMAE スコアを Baseline 比で**約 100 倍（0.004 → 0.394）**へと劇的に向上させた。特筆すべきは、VideoMAE の Top-1 予測クラスが、曖昧な ”dancing” ではなく、明確にターゲット動作である **”somersaulting”（宙返り）** に変化した点である。これは、提案手法が単にノイズを増やして偶然の動きを誘発したのではなく、モデルが潜在的に保持していた物理運動の知識を、適切な「緩和（Relaxation）」によって引き出すことに成功したことを証明している。我々は、大規模動画モデルにおける「Static Death」問題は、本手法によって克服可能であると結論付ける。

4.4.2 持続的緩和の必要性 ($p = 0.7$)

本実験において、減衰パラメータ p は 0.7（緩やかな減衰）が最適であった。これは、バク転のような「踏み切り→跳躍→回転→着地」という一連の複雑なシーケンスを生成するためには、初期の一瞬だ

けでなく、中盤にかけてもある程度の自由度（緩和状態）を持続させる必要があることを示唆している。$p = 2.0$ のような急減衰では、回転の途中でモデルの静的バイアスが復帰してしまい、動作が不完全になる（Robot Dancing 化する）ことが分かった。

以上の結果より、「初期の拘束を適度に、かつ持続的に緩める」という **Sustained Relaxation** 戦略こそが、高一貫性モデルに動的アクションを実行させるための鍵であることが明らかとなった。

5　まとめ

本研究では、大規模動画生成モデルの「Static Death」問題に対し、Adaptive CFG と Temporal Unlearning を用いた動的制御手法を提案した。実験の結果、初期推力（Beta）と急減衰（Power）、そして忘却（Blur）の 3 要素を適切に組み合わせることで、従来トレードオフの関係にあった「一貫性」と「可塑性」を高い次元で両立できることを実証した。これにより、FramePack 等の強力な慣性を持つモデルにおいても、意図した通りの動的アクション生成が可能となった。

参考文献

[1] HunyuanVideo Authors, et al. ”HunyuanVideo: A Large-scale Video Generation Model.” 2024.

[2] Wu, T., et al. ”FreeInit: Bridging Initialization and Inference for better Video Generation.” 2023.

[3] Hertz, A., et al. ”Prompt-to-Prompt Image Editing with Cross Attention Control.” 2022.