# 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

Adaptive Control of Temporal Inertia in Large-Scale Video Generation Models

著者名 *1
Author Name

*1所属
Affiliation

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) 生成初期に過剰な拘束を緩和する **Adaptive CFG (Relaxation)**、(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 Unlearning** 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 AI、特に Diffusion Transformer (DiT) ベースのモデル [1] は、近年目覚ましい進化を遂げている。数秒の動画生成であれば、全フレームを一度にメモリに展開し、相互の注意機構 (Self-Attention) によって高い一貫性を保つことが可能である。しかし、分単位あるいは無限の長尺動画を生成する場合、メモリ制約から「スライディングウィンドウ」や「自己回帰的 (Autoregressive)」な手法を取らざるを得ない。

FramePack 等の自己回帰モデルは、過去のフレームを条件として現在のフレームを生成することで、長時間の整合性を維持する。しかし、この強力な「過去への参照」は、諸刃の剣でもある。我々の予備実験において、これらのモデルは「消える (disappear)」「バク転する (backflip)」といった、状態の非連続な変化を伴うプロンプトを無視し、初期状態（立ち姿など）を維持し続ける傾向が確認された。我々はこの現象を、モデルが過去の文脈に過剰に囚われる **”Static Death”（静的死）** と呼ぶ。

本研究の目的は、この ”Static Death” を克服し、学習済みモデルのパラメータを変更することなく（Training-free）、推論時の介入のみで動的なアクション生成を実現することである。

## 2. 関連研究

### 2.1 大規模動画生成モデルと一貫性

Sora や HunyuanVideo [1] に代表される大規模モデルは、膨大なデータセットによる学習を通じて、物理法則や動作の事前知識を獲得している。しかし、生成時においては「時間的一貫性 (Temporal Consistency)」が最優先される傾向にあり、特に自己回帰的な生成においては、直前のフレームとの連続性が強く強制される。これは、背景の固定や人物のアイデンティティ維持には有効だが、急激なアクションの生成を阻害する要因となる。

### 2.2 推論時介入による制御

学習済みモデルに追加学習を行わずに制御を行う試みとして、FreeInit [2] や Adaptive Low-Pass Guidance (ALG) [4] がある。FreeInit は初期ノイズの再利用によって一貫性を向上させる手法であり、逆に言えば Static Death を助長する可能性がある。一方、ALG は条件画像から高周波成分を除去することで動きを促進する手法であり、本研究の方向性と近い。しかし、ALG はあくまで入力画像の前処理に留まっており、拡散モデルの生成ダイナミクスそのもの（CFG など）への介入は行なっていない。

### 2.3 拡散モデルの周波数特性と構造形成

拡散モデルの生成プロセスには、周波数領域における明確なバイアスが存在することが知られている。Choi ら [2] の分析によれば、デノイジングプロセスの初期段階（高ノイズ領域）では、画像の「低周波成分（大まかな構図・配置）」が決定され、終盤（低ノイズ領域）にかけて「高周波成分（細部・テクスチャ）」が形成される。既存の FreeInit 等の手法は、この特性を利用し、初期ステップの潜在変数を固定・反復することで、低周波成分（構造）の一貫性を強化しようとした。これは「歩行」のような定常的な動作には有効であるが、「バク転」のようなグローバルな構造変化（直立 $\rightarrow$ 倒立）を伴うアクションにおいては、致命的な制約となる。既存の構造バイアスが初期段階で固定されてしまうため、モデルは新たな姿勢へと遷移できなくなるのである。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。我々は、初期段階（低周波生成フェーズ）においてのみ、CFG や Attention による拘束を意図的に破壊（Relaxation）することで、グローバルな構造変化を許容する。提案手法における $\beta$ と $p$ は、まさにこの「低周波領域への介入強度と期間」を制御するパラメータと解釈できる。これにより、モデルは学習済みの物理知識（どう回転すべきか）を維持しつつ、初期状態の構造的制約（立ち姿）から解放されることが可能となる。

連絡先: 未定

## 3. 課題分析と仮説：なぜ Static Death は起きるのか

### 3.1 エネルギー地形による解釈

拡散モデルの生成過程は、エネルギーポテンシャルの斜面を下り、安定状態（極小値）へ収束するプロセスと見なせる。Static Death が発生している状態では、初期フレーム（立ち姿）の周辺に、極めて深く急峻なポテンシャルの谷（Deep Valley）が形成されていると考えられる。モデルはこの谷底に捕らわれており、プロンプトが「バク転」を指示しても、そのエネルギー障壁を越えて別の谷（バク転状態）へ遷移することができない。

### 3.2 初期仮説の失敗：Impulse (Boost) アプローチ

当初、我々はこの障壁を越えるためには「強い力」が必要であると考えた。すなわち、CFG スケールを負の値に設定したり（$\beta < 0$）、初期ノイズを増幅させることで、モデルを強制的に谷から押し出す **"Impulse" (Boost)** 戦略である。実際に $\beta = -1.0$（初期 CFG を強化）として予備実験を行ったところ、VideoMAE スコアは 0.007 と Baseline (0.004) に比して有意な改善が見られず、Top-1 予測クラスも "spinning poi" (7%) や "breakdancing" という脈絡のない結果となった。映像を詳細に分析すると、人物がバク転をするのではなく、手足が異常に伸長したり、多重露光のように身体が分裂したりする「構造的崩壊（Morphological Collapse）」が確認された。これは、過剰なエネルギー注入により、モデルがポテンシャルの谷を越えるどころか、地形そのものを破壊し、物理的に成立しないカオス状態（Spinning Poi のような意味不明な回転）へと発散してしまったことを意味する。この結果から、単に「力を加える」だけでは、繊細な構造変化（バク転）を誘導することは不可能であり、逆に「構造を壊してしまう」リスクが高いことが判明した。

### 3.3 修正仮説：Relaxation アプローチ

この失敗から、我々は「力で押すのではなく、壁を低くすればよいのではないか」という着想を得た。すなわち、生成の初期段階においてのみ、CFG スケールや Attention の制約を意図的に弱めることで、ポテンシャル地形全体を平坦化（Flattening）する **"Relaxation"** 戦略である。壁が低くなれば、モデルはわずかな駆動力（プロンプトの指示）でも容易に現在の谷を脱出し、隣接する目的の谷へと遷移できるはずである。本研究ではこの仮説に基づき、適応的なパラメータ制御手法を設計した。

## 4. 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。これらの手法は、先行研究で明らかにされた拡散モデルの特性に基づき設計されている。

### 4.1 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：

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

ここで $s$ は定数のガイダンススケールである。

#### 4.1.1 Relaxation Strength $\beta$: ALG からの拡張

先行研究である ALG [4] は、初期入力の高周波成分を落とすことで動きを促進できることを示した。これは「構造的制約を弱める」ことを意味する。我々はこの知見を拡張し、入力画像だけでなく、生成プロセスそのものの制約（CFG スケール）を初期段階で弱めることで、より直接的にモデルの探索空間を広げることができると考えた。具体的には、初期のガイダンススケールを減衰させる係数 $\beta$ を導入する。これにより、モデルはプロンプトの指示（動的ベクトル）に対してより敏感に反応できるようになる。

#### 4.1.2 Decay Power $p$: FreeInit の周波数特性に基づく制御

FreeInit [2] は、拡散プロセスの初期段階が画像の「低周波成分（大まかな構図）」を決定的に左右することを示した。逆に言えば、初期段階を過ぎれば構図は固定され、以降はディテールの生成に移行する。したがって、Relaxation（構造破壊）の効果は、構図が決定される「初期段階」のみに限定されるべきであり、後半まで持続するとディテールの崩壊を招く。この「介入の引き際」を厳密に制御するために、減衰パラメータ $p$ を導入した。提案する **Relaxation Curve** は以下の通りである：

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

各パラメータの定義と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。ALG の知見に基づき、初期の構造的制約を緩和する度合いを決定する。$\beta = 1.0$ のとき、初期スケールは $s_{min}$ まで低下し、探索自由度を最大化する。
- $p$ (**Decay Power**): FreeInit の知見に基づき、低周波生成フェーズ（初期）から高周波生成フェーズ（後半）への移行に合わせて緩和を終了させるための減衰指数。$p > 1$（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、後半の整合性を担保する。

### 4.2 Temporal Unlearning: Attention マップの拡散

Prompt-to-Prompt [3] は、生成される物体の形状や位置（アイデンティティ）が Self-Attention マップに強く依存していることを明らかにした。Static Death が起きている状態では、過去フレームの Attention マップ（立ち姿）が強固に維持されていると解釈できる。そこで本手法では、Self-Attention 層の Key ($K$) および Value ($V$) に対して時間軸方向の平滑化（Gaussian Blur）を適用する **Temporal Unlearning** を導入した。

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

これにより、Attention における時間的な局所性（過去への固執）を物理的に拡散させ、プロンプトによる新しい状態への遷移（Overwrite）を容易にする。これは ALG のような入力画像処理ではなく、モデル内部の記憶機構への直接介入である。

## 5. 実験・考察

### 5.1 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。

### 5.2 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。

### 5.3 定量的評価結果

### 5.4 定量的評価

#### 5.4.1 Relaxation Strength ($\beta$) の影響

初期ガイダンスの緩和強度 $\beta$ を、負の値（Boost）から正の値（Relaxation）まで変化させた結果を表 1 に示す。

表 1: Relaxation Strength ($\beta$) による性能比較 ($p = 0.7, \sigma_{blur} = 0.6$ 固定)

| Condition | $\beta$ | VideoMAE | Top Class | Prob. |
|---|---|---|---|---|
| Forced (Boost) | -1.0 | 0.007 | spinning poi | 0.07 |
| Forced (Boost) | -0.5 | 0.013 | breakdancing | 0.18 |
| Baseline | 0.0 | 0.004 | *Static (dancing)* | - |
| Ours (Mild) | 0.5 | 0.130 | somersaulting | 0.13 |
| **Ours (Best)** | **0.75** | **0.394** | **somersaulting** | **0.39** |
| Ours (Over) | 1.0 | 0.028 | capoeira | 0.09 |

表 1 が示すように、$\beta < 0$ の設定（初期 CFG を強める Boost 戦略）では、VideoMAE スコアは低迷し、生成されるクラスもターゲットとは無関係なもの（spinning poi など）となった。これは、初期の拘束を強めることが、逆にモデルを局所解（Local Minima）に閉じ込め、動的な遷移を阻害したためと考えられる。一方、$\beta > 0$（Relaxation 戦略）ではスコアが劇的に向上した。特に $\beta = 0.75$ において VideoMAE は **0.394** に達し、クラス分類も明確に "somersaulting" となった。しかし、$\beta = 1.0$ まで強めると逆にスコアが低下しており、過度な緩和は生成の崩壊を招くことが示唆された。$\beta = 0.75$ が「構造維持」と「動的遷移」の最適なトレードオフ点（Sweet Spot）であると結論付けられる。

### 5.5 定性的な失敗モードの分析

定量スコアだけでなく、生成された映像の挙動を詳細に観察することで、各パラメータが生成ダイナミクスに与える影響がより明確になった。以下に代表的な失敗モードを分類する。

#### 5.5.1 過剰な初期推力によるカオス化 (Chaos by Boost)

$\beta < 0$（Boost 設定）において、予想に反して VideoMAE スコアが極めて低かった（0.007 等）原因は、生成映像のカオス化にある。「バク転」という特定のアクションに遷移するのではなく、初期フレームの人物が意味不明な回転運動を始めたり（Spinning Poi）、手足が分裂・変形する（Morphological Collapse）現象が多発した。これは、エネルギー地形において、無理やりポテンシャルの壁を乗り越えさせようとした結果、バク転という「正解の谷」ではなく、物理的にあり得ない「異常な谷（Local Minima）」に落ち込んでしまったと解釈できる。力任せの介入（Impulse）では、繊細なアクション制御は不可能であることが示された。

#### 5.5.2 急減衰による空中静止 (Mid-air Freeze)

一方、Relaxation そのものは成功しているが、減衰率 $p$ が大きすぎる（$p = 2.1$）場合に見られたのが「空中静止」現象である。生成初期にはスムーズに跳躍を開始するが、体が空中に浮いた瞬間に Relaxation 効果が切れ（スケールが強拘束に戻り）、その姿勢のまま静止あるいは元の立ち姿勢に無理やり戻ろうとする挙動（Elastic Snap-back）が確認された。バク転のような、離陸から着地まで約 1〜2 秒を要するアクションにおいては、その全工程を通じて適度な可塑性（Plasticity）を維持し続ける「持続的な緩和 ($p \approx 0.7$)」が不可欠である。

#### 5.5.3 Decay Power ($p$) による時間制御

緩和の効果を時間的にどう減衰させるかを決定する $p$ の影響を表 2 に示す。ここでは過剰緩和気味であった $\beta = 1.0$ をベースに検証した。

表 2: Decay Power ($p$) の影響 ($\beta = 1.0, \sigma_{blur} = 0.6$)

| Condition | $p$ | VideoMAE | Top Class |
|---|---|---|---|
| Ours (Sustained) | 0.7 | 0.028 | capoeira |
| Ours (Linear) | 1.0 | 0.026 | capoeira |
| Ours (Rapid) | 2.1 | 0.019 | capoeira |

$p = 2.1$（急減衰）の場合、初期の緩和効果が急速に失われるため VideoMAE スコアは最も低くなった。バク転のような滞空時間の長いアクション（約 1 秒〜2 秒）を生成する場合、初期数ステップだけでなく、中盤にかけても緩和効果を持続させる（$p < 1.0$）ことが重要であることがわかる。

### 5.6 理論的考察：エネルギー地形モデル

なぜ初期の緩和（Relaxation）が必要なのか。これを「エネルギー地形（Energy Landscape）」の観点から考察する。拡散モデルの生成過程は、エネルギーポテンシャルの斜面を下るプロセスに例えられる。Static Death が発生するモデルでは、初期画像（立ち姿）の周辺に深く急峻なポテンシャルの谷（Deep Valley）が形成されていると考えられる。通常の CFG（$\beta = 0$）や Boost（$\beta < 0$）は、この谷底へ向かう力を強めるため、モデルは谷から脱出できず、結果として動きのない映像が生成される。対して本手法の Relaxation（$\beta > 0$）は、一時的にポテンシャルの勾配を平坦化（Flattening）する効果を持つ。これにより、モデルの状態は初期値の谷から容易に脱出し、隣接する別の谷（バク転している状態）へと遷移することが可能になる。ただし、平坦化させすぎると（$\beta = 1.0$）、モデルはどの谷にも収束できず、意味のないノイズ（Chaos）へと発散してしまう。これが $\beta = 1.0$ でスコアが悪化した理由である。適度な $\beta = 0.75$ だけが、初期の谷からの脱出と、目的の谷への収束を両立させることができるのである。

## 6. 議論

### 6.1 安定性と可塑性のトレードオフ

本研究の結果は、動画生成モデルにおける「安定性（Stability）」と「可塑性（Plasticity）」の根源的なトレードオフを浮き彫りにした。HunyuanVideo のような基盤モデルは、学習データの大規模化により極めて高い安定性を獲得したが、それは同時に可塑性の喪失（Static Death）を招いた。本手法の Relaxation は、推論時にこのバランスを動的に操作する「ツマミ」を提供するものである。特に、$\beta = 0.75$ という値が最適であった事実は、完全に拘束を解く（$\beta = 1.0$）のでなく、「ある程度の安定性を保ちつつ、可塑性を注入する」という微妙なバランス制御が重要であることを示唆している。

### 6.2 長尺動画生成への応用

本手法は、単発のアクション生成に留まらず、長尺なストーリー動画の生成に対しても強力なツールとなり得る。例えば、映画の脚本において「静かな会話シーン」と「激しい戦闘シーン」が混在する場合、従来のモデルでは常に一定のパラメータで生成せざるを得なかった。しかし本手法を用いれば、シーンのメタデータ（激しさ）に応じて $\beta$ や $p$ を動的に変化させることで、静的なシーンでは高画質・高一貫性を維持し（Relaxation OFF）、動的なシーンでは大胆なカメラワークやアクションを許容する（Relaxation ON）といった、演出意図の反映（Directorial Control）が可能になる。これは、「生成 AI による映画制作」の実現に向けた重要な一歩である。

### 6.3 限界と今後の課題

現状の課題として、最適な $\beta$ や $p$ の値がプロンプトの種類（動きの激しさ）に依存する点が挙げられる。「歩く」程度の動きであれば弱い緩和で十分かもしれないが、「爆発」のようなシーンではより強い緩和が必要になる可能性がある。プロンプトの内容から最適なパラメータを推定する「Meta-Controller」の開発が今後の課題である。

### 6.4 既存手法との比較

動画生成の制御手法としては ControlNet や MotionLora などが存在するが、これらは追加の学習（Training）や外部データセットを必要とする。対して本手法は、推論時のハイパーパラメータ（$\beta, p, \sigma_{blur}$）の調整のみで実現可能（Training-free）であり、計算コストおよび実装コストにおいて圧倒的な優位性を持つ。特に大規模な基盤モデル（Foundation Model）に対して、再学習なしで動的な挙動を修正できる点は、実用上極めて価値が高い。

## 7. 結論

本研究では、大規模動画生成モデルにおける深刻な課題 "Static Death" を定義し、その克服に向けた推論時介入手法を提案した。本研究の主な貢献は以下の 3 点に集約される。

1. **Static Death の現象論的解明**: モデルが過去の文脈に過剰に囚われる現象を「エネルギー地形における深い局所解への埋没」としてモデル化し、従来のインパルス（Boost）アプローチがカオスを招くことを明らかにした。

2. **Relaxation 理論の確立**: 力で脱出するのではなく、初期拘束を一時的に緩和して障壁を下げる「Relaxation」アプローチを提唱し、その有効性を理論的・実験的に実証した。

3. **劇的な性能向上**: 提案手法により、追加学習なしで VideoMAE スコアを約 100 倍（0.004 → 0.394）に向上させ、従来不可能であった「バク転」等の動的アクション生成を実現した。

本手法は、高画質化・安定化が進む今後の動画生成モデルにおいて、失われがちな「動的な表現力」を取り戻すための標準的な制御技術となることが期待される。

## 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.

[4] Gu, Y., et al. "Adaptive Low-Pass Guidance for Image-to-Video Generation." 2024.

[5] Research Team, "Stable Video Infinity." 2024.

[6] Research Team, "MotionRAG: Retrieval-Augmented Motion Generation." 2024.