一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report

Dual Attention機構と動画拡散モデルを活用した長尺動画のカラー化

木村倫太郎† 　柳井　啓司†

† 電気通信大学
E-mail: †kimura-r@mm.inf.uec.ac.jp, ††yanai@cs.uec.ac.jp

あらまし　長尺動画のカラー化における自然な発色と時間的一貫性、計算量の課題に対し、テキストプロンプトで色制御が可能な新たな手法を提案する。本手法は、サンプリングしたアンカーフレームを Dual Attention を導入した Stable Diffusion で個別にカラー化し、中間フレームを双方向からの推論と ControlNet で制御された Stable Video Diffusion で補間して生成する。実験では DAVIS30 および Videvo データセットにおいて、主要指標で最先端手法を上回る性能を達成した。本手法は、高い知覚品質と時間的安定性を実現しつつ、一定のメモリ消費量で長尺動画を効率的に処理可能である。

キーワード　カラー化、拡散モデル、動画カラー化、テキストプロンプトによる色制御

Long Video Colorization with Dual Attention and Video Diffusion Model

Rintaro KIMURA† and Keiji YANAI†

† The University of Electro-Communications
E-mail: †kimura-r@mm.inf.uec.ac.jp, ††yanai@cs.uec.ac.jp

Abstract　We propose a novel method capable of color control via text prompts to address the challenges of natural coloration, temporal consistency, and computational complexity in long video colorization. In this method, sampled anchor frames are individually colorized using Stable Diffusion incorporating Dual Attention, and intermediate frames are generated by interpolation using Stable Video Diffusion controlled by bidirectional inference and ControlNet. Experiments on the DAVIS30 and Videvo datasets demonstrated that our method outperforms state-of-the-art methods in key metrics. Our approach enables efficient processing of long videos with constant memory consumption while achieving high perceptual quality and temporal stability.

Key words　colorization, diffusion models, video colorization, text-guided colorization

1. はじめに

歴史的資料や記録映像はモノクロのまま残されており、これらをカラー化して保存・活用したいという需要は今もなお多く存在する。近年では深層学習の進展により、自動あるいは半自動でのカラー化が活発化しているが、時間的な一貫性と自然な発色の両立は主要な課題であり、特に拡散モデルを用いた既存手法では、フレーム数の増加に伴う計算量の増大により、長尺動画の処理が困難であった。

そこで本研究では、テキストプロンプトによる柔軟な色制御と、長尺動画の一貫したカラー化を両立する新たなフレームワークを提案する。本手法は、大規模に事前学習された拡散ベースの画像・動画生成モデルを活用し、高い知覚品質と時間的安定性を両立したカラー化を実現しつつ、安定して処理が可能である。

2. 関連研究

2.1 動画カラー化手法の種類と課題

静止画のカラー化技術は、CNN や GAN、そして拡散モデルの登場により飛躍的に向上した。これを動画へ拡張する際、最大の課題となるのが時間的一貫性の確保である。

従来の手法は主に 5 つに大別される。

- **オプティカルフローベース**：隣接するフレーム間の画素や特徴の動き、オプティカルフローを計算し、それに基づいて色情報を伝播させることで時間的な一貫性を保つ手法。
- **スクリブルベース**：ユーザーが動画フレーム内の特定の領域に色を指定し、それをターゲットとなるオブジェクト全体や後続のフレームへと伝播させる手法。
- **参照ベース**：色付きの参照画像から対応する色を抽出し、それをグレースケールの動画フレームに転送するアプローチをとる手法。

- **全自動**：参照画像やユーザーによるスクリブルの入力を必要とせず、エンドツーエンドでカラー化する手法。
- **プロンプトベース**：テキストプロンプトにより、自然言語で自在に色を制御しながらカラー化する手法。

2.2 本研究の位置づけと基盤技術

本研究は、テキストプロンプトによって動画の色調を制御するプロンプトベースのアプローチに位置づけられる。しかし、各フレームを個別にテキストから生成する従来の手法とは異なり、本手法はテキストプロンプトによってカラー化された中央フレームを参照画像として扱い、その色情報を後続のフレームへと伝播させるハイブリッドな構成を採用した。

さらに、本手法はユーザーが具体的な色指定を行いたい場合には詳細なプロンプトを受け付ける一方で、汎用的なプロンプトをデフォルトとして設定することで、ユーザー入力を一切必要としない全自動手法としても機能する。

基盤技術として、参照画像の生成には、テキスト理解と画像生成の整合性に優れた Qwen-Image-Edit [1] を採用する。色情報の転写には、Kosugi [2] が提案した Dual Attention 機構を導入し、参照画像の色情報を入力画像の構造に合わせて高精度に注入する。さらに、中間フレームの補完には動画生成モデル Stable Video Diffusion (SVD) [3] を利用する。ここでは Liu ら [4] の枠組みに基づき、始点・終点からの双方向推論と、エッジやオプティカルフローを用いたマルチモーダル ControlNet による構造制御を組み合わせる。これにより、拡散モデル特有のドリフト現象を抑制し、長尺動画においても高品質かつ時間的に滑らかなカラー化を実現する。

3. 提 案 手 法

本論文では、グレースケール動画とテキストプロンプトを入力とし、時間的一貫性と高い知覚品質を両立する新たな動画カラー化フレームワークを提案する。

3.1 パイプライン概要

提案手法の全体パイプラインを図 1 に示す。本手法は、処理の役割に応じて、参照画像生成、アンカーフレームのカラー化、中間フレーム補完、LAB 空間での統合の四つの段階から構成される。

第一段階では、カラー化の基準となる画像を生成するため、入力動画の中央フレームとテキストプロンプトを用いて、画像編集モデル Qwen-Image-Edit によるカラー化を行う。これにより得られた画像は、アンカーフレームのカラー化において、色彩のスタイルとトーンを規定する参照画像として機能する。

第二段階では、動画全体から一定間隔でアンカーフレームを抽出し、第一段階で生成された参照画像を用いて各アンカーフレームを高精度にカラー化する。

ここでは、小杉の手法 [2] に基づき、Stable Diffusion 1.4 (SD1.4) をベースとした拡散モデルに対して Dual Attention 機構を導入した潜在空間上での色転写を行う。これにより、全アンカーフレームに対し、参照画像と意味的に一致するオブジェクトを統一された色味でカラー化することができる。

第三段階では、カラー化されたアンカーフレーム間を埋める中間フレームの生成を行う。ここでは、Liu ら [4] の手法に基づき、動画拡散モデルである Stable Video Diffusion (SVD) を拡張した双方向生成アプローチを採用する。隣接する 2 枚のアンカーフレームをそれぞれ始点・終点の条件として前向きおよび後向きの推論を行い、元のグレースケール動画から抽出した Canny エッジとオプティカルフローを ControlNet [5] で注入する。これにより両端のアンカーフレームと整合し、かつ被写体の動きに忠実な遷移を実現する。

最終段階では、生成されたカラー動画の色差成分と、元の入力グレースケール動画の輝度成分を LAB 色空間上で合成する。この処理により、第三段階で生成したカラー動画のボケや構造的なアーティファクトを排除し、元動画が持つ本来の解像感とテクスチャを忠実に保持したカラー動画を得る。

3.2 参照画像生成

動画全体で一貫したカラー化を実現するため、本手法では入力グレースケール動画から中央フレーム v_{mid} を抽出し、これを基準画像としてテキスト条件付きでカラー化する。生成モデルには Qwen-Image-Edit [1](注1) を用いる。Qwen2.5-VL を条件エンコーダ、MMDiT を拡散バックボーンとする構成により、入力構造を保ちながらテキスト指示に忠実な色編集が可能である。具体的には、v_{mid} と色指定を含むプロンプト T を入力し、構造を維持したカラー参照画像 I_{ref} を生成する。得られた I_{ref} は、後段のアンカーフレームカラー化における唯一の色参照として機能し、プロンプト意図に沿った一貫した色伝播を支える。

3.3 アンカーフレームのカラー化

このプロセスでは、Stable Diffusion 1.4 (SD1.4)(注2) をベースモデルとし、小杉 [2] が提案した Dual Attention 機構を高解像度な潜在空間へ適用することで、参照画像の色彩を入力フレームの構造に合わせて正確に転写する。

処理は、拡散過程の反転による特徴抽出、Dual Attention を用いた色転写、そして統計的補正による洗練の 3 つのステップから構成される。

3.3.1 拡散過程の反転と特徴抽出

まず、入力されるアンカーフレーム（グレースケール）と参照画像（カラー）の前処理を行う。SD1.4 の VAE の仕様に合わせ、入力画像は解像度が 64 の倍数となるようにリサイズおよびパディング処理を施す。

続いて、入力画像、参照画像、および参照画像の輝度成分のみを抽出したグレースケール参照画像の 3 枚を、それぞれ VAE エンコーダを用いて潜在表現 z_0^{in}、z_0^{ref}、z_0^{refL} へと変換する。

次に、これらの潜在表現に対して DDIM Inversion を適用し、拡散過程を遡ることで、各画像の構造情報を保持した初期ノイズ z_T を推定する。この反転プロセスの各タイムステップ t において、U-Net 内の Self-Attention 層から以下の特徴量（Query: Q, Key: K, Value: V）を抽出し、保存する。

- 入力画像（z^{in}）からは、Q_t^{in} および K_t^{in} を保存する。これらは入力画像の形状と構造情報を保持している。

(注1)：https://huggingface.co/Qwen/Qwen-Image-Edit-2509
(注2)：https://huggingface.co/CompVis/stable-diffusion-v1-4

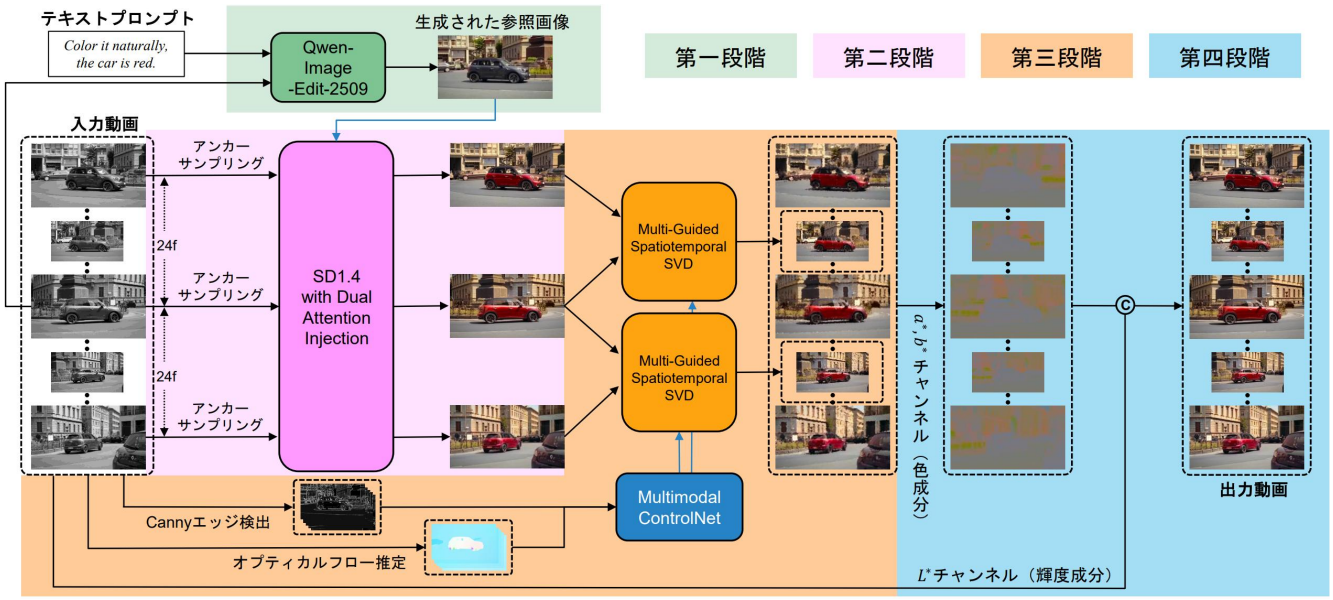

図 1　パイプライン全体図

- 参照画像（z^{ref}）からは、K_t^{ref} および V_t^{ref} を保存する。これらは参照画像の色情報とテクスチャ情報を保持している。
- グレースケール参照画像（z^{refL}）からは、K_t^{refL} を保存する。これは参照画像の構造的な対応関係を計算するために用いられる。

これらの保存された特徴量は、次段の生成プロセスにおいて、異なる画像間でのアテンション計算に利用される。

3.3.2　Dual Attention 機構

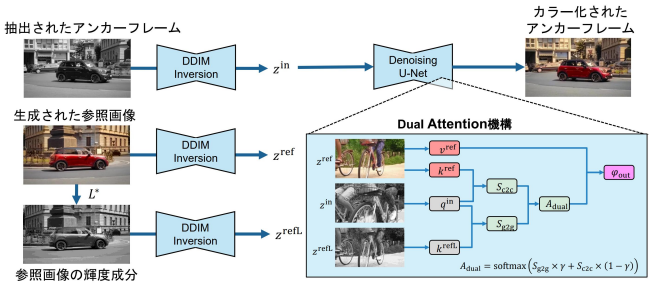

図 2　Dual Attention 機構の概略図

カラー化されたアンカーフレームの生成は、DDIM サンプリングによってノイズを除去していくことで行われる。

初期潜在変数 z_T には、AdaIN (Adaptive Instance Normalization) [6] を適用し、その平均と分散を参照画像の特徴に合わせて調整することで、色生成の初期状態を整える。

拡散モデルの更新ステップでは、通常の Self-Attention 層を図 2 のように Dual Attention 機構に置き換えて推論を行う。ここでは、保存された特徴量を用いて 2 つのアテンションマップを作成し、それらを統合することで、構造の維持と色の転写を同時に達成する。

第一に、入力画像の構造と参照画像の構造の対応を取るため、保存された入力のクエリ q_t^{in} とグレースケール参照のキー k_t^{refL} を用いて Gray-to-Gray Attention を計算する。

$$S_{g2g} = \frac{q_t^{in}(k_t^{refL})^T}{\sqrt{d}} \tag{1}$$

第二に、現在の生成過程におけるクエリ q_t^{out} と参照のキー k_t^{ref} を用いて Colorized-to-Color Attention を計算する。

$$S_{c2c} = \frac{q_t^{out}(k_t^{ref})^T}{\sqrt{d}} \tag{2}$$

これらを重み付け混合し、最終的なアテンションマップ A_{dual} を得る。

$$A_{dual} = \text{softmax}(S_{g2g} \times \gamma + S_{c2c} \times (1-\gamma)) \tag{3}$$

このマップに基づいて参照画像のバリュー v_t^{ref} を加重平均し、入力画像の特徴量へ注入することで、意味的に対応する領域への色転写を行う。

さらに、ノイズ予測においては、Classifier-free Colorization Guidance を適用する。Dual Attention による色転写を行う条件付き推論と、行わない無条件推論の双方を実行し、その差分を増幅させて更新を行うことで、入力の構造を保ちつつ、参照画像の色鮮やかさを反映した画像を生成する。

3.4　中間フレーム補間

前段階で生成されたカラー化アンカーフレームは時間的に疎な状態であるため、これらを境界条件として、その間のフレームを高精度に生成し、時間的に滑らかなカラー動画を再構成する必要がある。本手法では、Liu ら [4] が提案した AnchorSync のマルチモーダルガイド付き補完の枠組みを採用し、Stable Video Diffusion (SVD)(注3) を用いた条件付き動画生成を行う。

(注3)：https://huggingface.co/stabilityai/stable-video-diffusion-img2vid-xt

3.4.1 構造・動作制約の抽出

単なる映像補間では、元動画の中間フレームと動きや構造が異なる画像が出力される可能性がある。そこで、元のグレースケール動画が持つ形状や動きを完全に保存したまま色彩のみを補完するために、マルチモーダル ControlNet [4] を用いた構造制御を導入する。

まず、各フレームに対して Canny エッジ検出とオプティカルフロー推定を行う。これらの条件入力を処理するために、Liu ら [4] によって追加学習されたマルチモーダル ControlNet を採用する。

抽出された特徴量は、凍結された SVD の U-Net の各ブロックへ注入され、元のグレースケール動画のダイナミクスを忠実に再現しつつ、アンカーフレームから指定された色情報を適切に伝播させることを可能にする。

3.4.2 動画拡散モデルによる双方向推論

中間フレームの生成には、画像条件付き動画生成モデルである SVD を基盤として用いる。一般的な SVD は、開始フレームのみを条件として時間方向に順次生成を行うため、生成シーケンスが長くなるにつれて誤差が累積し、終点のアンカーフレームと不連続になるドリフト現象が発生するという課題がある [4]。

本手法では、この問題を解決するために双方向時間フレーム融合を導入する。具体的には、隣接する 2 枚のアンカーフレーム f_i'（始点）と f_{i+1}'（終点）を用い、以下の 2 つの生成プロセスを並行して実行する。

（1） **順方向生成 (Forward Pass):** 始点 f_i' を条件画像として入力し、時間順方向に拡散モデルの推論を行う。これにより得られる各タイムステップの潜在表現を $Z^{forward}$ とする。

（2） **逆方向生成 (Backward Pass):** 終点 f_{i+1}' を条件画像として入力し、時間逆方向（または逆再生したシーケンスに対する推論）を行う。これにより得られる潜在表現を $Z^{reverse}$ とする。

各デノイジングステップにおいて、これら 2 つの潜在表現を線形合成し、融合された潜在表現 Z_j^{fuse} を得る。

$$Z_j^{fuse} = \alpha \cdot Z_j^{forward} + (1-\alpha) \cdot Z_j^{reverse} \tag{4}$$

この融合処理により、生成される中間フレーム列は、始点付近では始点アンカーの特徴を、終点付近では終点アンカーの特徴を強く反映することになる。結果として、両端のアンカーフレームと整合性が取れた、アーティファクトの現れにくい滑らかな遷移を持つカラー動画区間が生成される。この処理をすべてのアンカー区間に対して適用し、連結することで、動画全体のカラー化を完了する。

3.5 LAB 空間での統合

中間フレーム補間により時間的に整合したカラー動画は得られるが、拡散モデルの生成過程ではテクスチャの低下や軽微なアーティファクトが生じることがある。一方、入力グレースケール動画は解像度と高周波成分を保持しており、構造情報の信頼性が高い。そこで本手法では、生成動画の色情報と入力動画の構造情報を分離し、再結合することで最終品質を向上させる。

具体的には、生成動画シーケンス V_{gen} と入力グレースケール動画シーケンス V_{in} を LAB 空間へ変換し、V_{gen} から色差成分 a^*, b^*、V_{in} から輝度成分 L^* を抽出する。その後、$L^*(V_{in})$ と $a^*, b^*(V_{gen})$ を統合して LAB 画像を再構成し、RGB へ逆変換して最終動画 V_{final} を得る。

この輝度置換により、色の一貫性を保ったまま、拡散生成で失われやすい細部構造を回復し、高品質なカラー化結果を実現する。

4. 実　　験

4.1 実装詳細

本手法の各段階におけるモデル構成とハイパーパラメータの詳細は以下の通りである。

4.1.1 モデル構成

第一段階の参照画像生成には、指示追従能力に優れた Qwen/Qwen-Image-Edit-2509 を使用した。第二段階のアンカーフレームのカラー化には、CompVis/stable-diffusion-v1-4 (SD1.4) をベースモデルとして採用し、計算精度は FP16 とした。第三段階の中間フレーム補完には、stabilityai/stable-video-diffusion-img2vid-xt (SVD) を使用し、構造制御には Panda-70M データセット [7] で学習されたマルチモーダル ControlNet [5] を用いた。オプティカルフローの推定には UniMatch (Unimatch-Debloated/gmflow-scale2) [8] を採用した。

4.1.2 ハイパーパラメータ設定

a） 参照画像生成:

入力動画の解像度を維持したまま、サンプリングステップ数 30、乱数シード 42 で推論を行った。

b） アンカーフレームのカラー化:

アンカーフレームの間隔は [4] に基づき、$K = 24$ とした。拡散過程には DDIM サンプラーを使用し、タイムステップ数は $T = 50$、$\eta = 0.0$ に設定した。Dual Attention 機構におけるアテンションマップの混合重みは、$\gamma_1 = 0.5$、$\gamma_2 = 1.0$ とし、Classifier-free Colorization Guidance のスケールは $w = 10.0$ とした。また、品質向上のための反復的洗練は 3 回実施し、初期潜在変数の AdaIN [6] 適用を有効化した。

c） 中間フレームの補完:

SVD への入力画像の解像度は、576×320（幅×高さ）にリサイズおよびセンタークロップを行って調整した。推論ステップ数は 25、モーションバケット ID は 127 に設定した。補完時のガイダンススケールは、始点から終点にかけて 1.0 から 8.0 へ線形に増加させ、ControlNet の適用強度は 1.0 とした。特徴注入の閾値は、Self-Attention 層に対して $\tau_{attn} = 0.44$、ResNet 層に対して $\tau_f = 0.65$ とした。

4.2 実験設定

4.2.1 データセット

提案手法の有効性を検証するため、動画カラー化のベンチマークとして広く用いられている DAVIS30 データセット [9] を使用した。提案手法のカラー化には、テキストプロンプトとし

て共通の *“Colorize it naturally with colorful tone.”* を使用した。

4.2.2 評価指標

生成されたカラー動画の品質を多角的に評価するため、以下の 5 つの指標を採用した。

- **PSNR (Peak Signal-to-Noise Ratio)** および **SSIM (Structural Similarity)**: 画素レベルおよび構造レベルでの信号忠実度を評価する。数値が高いほど正解に近いことを示す。
- **LPIPS (Learned Perceptual Image Patch Similarity) [10]**: 人間の視覚特性に基づいた知覚的な類似度を評価する。数値が低いほど自然であることを示す。
- **FID (Fréchet Inception Distance) [11]**: 生成画像分布と実画像分布の距離を測定し、画像のリアリズムを評価する。数値が低いほど高品質であることを示す。
- **CDC (Color Distribution Consistency) [12]**: 時間方向の色分布の安定性を評価する指標であり、値が小さいほどフリッカーが少ないことを示す。
- **Colorfulness [13]**: 生成結果の色彩の豊かさを定量化する指標であり、値が大きいほど色彩が豊かであることを示す。

4.2.3 比較手法

提案手法の性能を位置づけるため、以下の最先端手法との比較を行った。従来手法として、自動カラー化の AutoColor [14]、参照ベースの DeepExemplar [15] および DeepRemaster [16]、時間的整合性に特化した TCVC [12]、GAN ベースの VCGAN [17] を採用した。また、拡散モデルおよびプロンプトベースの最新手法として、ColorDiffuser [18]、L-C4 [19]、VanGogh [20] とも比較を行った。

4.3 定量的評価

DAVIS30 における定量的評価の結果を表 1 に示す。ColorDiffuser [18]、L-C4 [19] および VanGogh [20] の値はそれぞれの論文から、その他の比較手法の値は [18] から引用した。「↑」の書かれた指標は高いほど良く、「↓」の書かれた指標は低いほど良い性能を示す。**太字**は最良、<u>下線</u>は次点の結果を表す。なお、「-」のデータは未報告かつコード非公開で再現が不可能なデータである。

4.3.1 総合評価

提案手法は、DAVIS30 と Videvo の両データセットで PSNR・SSIM・LPIPS・FID の主要指標において高水準の性能を示し、特に LPIPS は DAVIS30 で 0.117、Videvo で 0.107 と比較手法を大きく上回った。Colorfulness も DAVIS30 で 43.51、Videvo で 38.98 を記録し、鮮やかさと自然さの両立を確認できた。一方、時間的一貫性（CDC）は一貫性特化法に及ばない場面があるものの、AnchorSync による双方向補完と色伝播により、長尺動画でも実用上十分な安定性を維持している。

4.4 定性的評価

図 3 に、自転車に乗る人物の動画を用いた従来手法との比較結果を示す。AutoColor [14] や VCGAN [17] などの手法では全体的に彩度が低く色が滲む傾向があり、DeepRemaster [16] では時間の経過とともに色の一貫性が失われている。一方、提案手法（Ours）は、Dual Attention 機構により参照画像の色情報を正確に反映しており、アンカーフレーム間での色の一貫性が保たれている。また、中間フレームにおいても ControlNet による構造制約が効いており、動きのあるシーンでもエッジの破綻なく滑らかなカラー化を実現している。

表 1　DAVIS30 における定量的評価結果。

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	CDC ↓	Colorfulness ↑
AutoColor [14]	24.41	0.915	0.264	83.05	0.003734	14.14
DeepExemplar [15]	21.78	0.846	0.325	77.26	0.004006	28.82
DeepRemaster [16]	21.95	0.848	0.354	97.54	0.005098	25.66
TCVC [12]	25.17	0.921	0.239	74.94	0.003649	21.72
VCGAN [17]	23.90	0.910	0.247	70.29	0.005303	15.89
ColorDiffuser [18]	23.73	<u>0.939</u>	0.213	<u>69.51</u>	<u>0.003607</u>	29.13
L-C4 [19]	<u>25.69</u>	0.933	0.209	–	**0.003114**	29.33
VanGogh [20]	23.20	<u>0.939</u>	<u>0.191</u>	–	–	**60.09**
Ours	**26.28**	**0.942**	**0.117**	**61.50**	0.003965	<u>43.51</u>

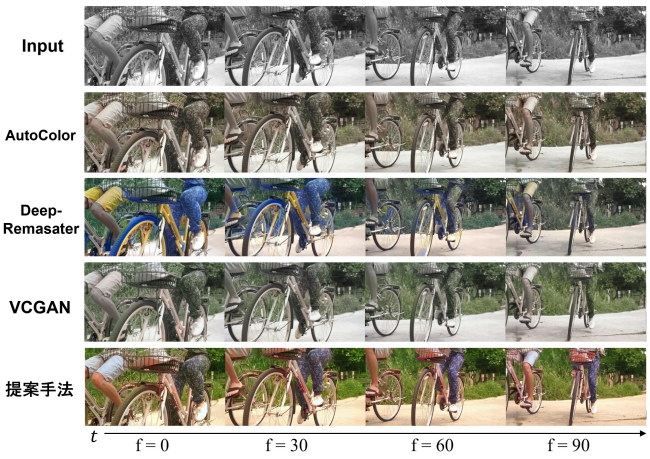

図 3　従来手法と提案手法によるカラー化結果の比較。

4.4.1 プロンプトによる制御性

図 4 は、同一のグレースケール動画に対して異なるテキストプロンプトを与えた場合の結果である。上段は「自然なカラー化」のみを指示した場合、下段は「男性のスーツは青、女性のシャツは黄色」と具体的に指定した場合である。提案手法は、第一段階の参照画像生成において高い言語理解能力を持つモデルを使用しているため、プロンプトを変更するだけで、ユーザーの意図に応じた具体的な配色を動画全体へ一貫して適用可能であることが実証された。

Prompt: *“Colorize it naturally with colorful tone.”*

Prompt: *“Colorize it naturally with colorful tone.*
*The man’s suit is **blue**, and the woman’s shirt is **yellow**”*

図 4　異なるプロンプトによる生成結果の比較。

4.4.2 長尺動画への適用

図 5 は、300 フレームを超える長尺動画を用いて検証を行った結果である。従来の動画カラー化における課題は、動画長に

比例して計算コストが増大することであった。これに対し提案手法は、動画全体をアンカー区間ごとに分割して処理するパイプラインを採用しているため、メモリ消費量を一定水準（約22GB）に抑えつつ処理が可能である。結果として、長時間のシーケンス全体を通しても色調のブレや破綻がなく、安定したカラー化が実現されている。

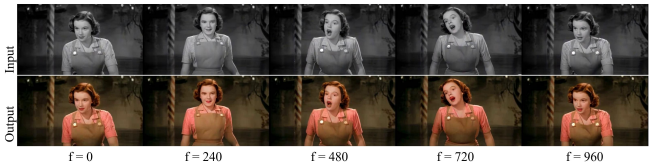

図 5　長尺動画のカラー化結果。

5. おわりに

5.1 まとめ

本研究では、テキストプロンプトによる色制御と時間的・空間的整合性を両立する動画カラー化手法を提案した。本手法は、参照画像生成、アンカーフレームカラー化、中間フレーム補完、LAB 空間統合の四段階で構成される。

有効性は次の 3 点で確認された。第一に、PSNR・SSIM に加え LPIPS・FID でも高い性能を示し、生成品質と構造保存を両立した。第二に、SVD で問題となるドリフトやフリッカーに対し、始点・終点アンカーからの双方向推論で時間的一貫性を改善した。第三に、Qwen-Image-Edit により言語指示だけで色調制御が可能となり、アンカー区間分割処理により 300 フレーム超の長尺動画でもメモリを一定に保って処理できることを示した。

5.2 今後の課題と展望

今後の課題は主に 4 点である。第一に、SVD の解像度・アスペクト比制約によりクロップやリサイズが必要となるため、可変比率対応モデルや解像度に依存しないパイプラインが必要である。第二に、中央フレーム 1 枚への参照依存により、遮蔽領域や画面外物体で色付け精度が低下するため、複数参照フレームの動的選択が有望である。第三に、激しい動きや多物体シーンではフロー誤差により追従が崩れるため、アンカー間隔の適応化や高精度トラッキング統合が必要である。第四に、現状は単一カット前提であり、シーンチェンジを含む長編動画では一貫性維持が難しい。今後はシーン検出とシーン単位の参照生成を組み合わせ、動画全体のトーンを統合的に制御する枠組みへ拡張する。

文　献

[1] C. Wu, J. Li, J. Zhou, J. Lin, K. Gao, K. Yan, S. mingYin, S. Bai, X. Xu, Y. Chen, Y. Chen, Z. Tang, Z. Zhang, Z. Wang, A. Yang, B. Yu, C. Cheng, D. Liu, D. Li, H. Zhang, H. Meng, H. Wei, J. Ni, K. Chen, K. Cao, L. Peng, L. Qu, M. Wu, P. Wang, S. Yu, T. Wen, W. Feng, X. Xu, Y. Wang, Y. Zhang, Y. Zhu, Y. Wu, Y. Cai, and Z. Liu, "Qwen-Image technical report," 2025. arXiv:2508.02324.

[2] S. Kosugi, "Leveraging the powerful attention of a pre-trained diffusion model for exemplar-based image colorization," IEEE Transactions on Circuits and Systems for Video Technology, vol.35, no.10, pp.10059–10069, 2025.

[3] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts, V. Jampani, and R. Rombach, "Stable video diffusion: Scaling latent video diffusion models to large datasets," 2023. arXiv:2311.15127.

[4] Z. Liu, Y. Wang, T. Wei, and C. Ma, "AnchorSync: Global consistency optimization for long video editing," Proc. of ACM International Conference Multimedia, pp.4494–4503, 2025.

[5] L. Zhang and M. Agrawala, "Adding conditional control to text-to-image diffusion models," Proc. of IEEE International Conference on Computer Vision, pp.3813–3824, 2023.

[6] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," Proc. of IEEE International Conference on Computer Vision, pp.1501–1510, 2017.

[7] T.-S. Chen, A. Siarohin, W. Menapace, E. Deyneka, H.-w. Chao, B.E. Jeon, Y. Fang, H.-Y. Lee, J. Ren, M.-H. Yang, and S. Tulyakov, "Panda-70M: Captioning 70m videos with multiple cross-modality teachers," Proc. of IEEE Computer Vision and Pattern Recognition, pp.13320–13331, 2024. https://github.com/snap-research/Panda-70M

[8] H. Xu, J. Zhang, J. Cai, J. Yang, H. Liu, Y. Zhang, and X. Tong, "Unifying flow, stereo and depth estimation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.45, no.11, pp.13941–13958, 2023.

[9] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," Proc. of IEEE Computer Vision and Pattern Recognition, pp.724–732, 2016.

[10] R. Zhang, P. Isola, A.A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," Proc. of IEEE Computer Vision and Pattern Recognition, pp.586–595, 2018.

[11] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," Advances in Neural Information Processing Systems, vol.30, pp.6626–6637, 2017.

[12] Y.-H. Liu, H.-Y. Zhao, K.C. Chan, X.-T. Wang, C.C. Loy, Y. Qiao, and C. Dong, "Temporally consistent video colorization with deep feature propagation and self-regularization learning," Computational Visual Media, vol.10, no.2, pp.375–395, 2024.

[13] D. Hasler and S.E. Suesstrunk, "Measuring colorfulness in natural images," Human Vision and Electronic Imaging VIII, vol.5007, pp.87–95, SPIE, 2003.

[14] C. Lei and Q. Chen, "Fully automatic video colorization with self-regularization and diversity," Proc. of IEEE Computer Vision and Pattern Recognition, pp.3748–3756, 2019.

[15] B. Zhang, M. He, J. Liao, P.V. Sander, L. Yuan, A. Bermak, and D. Chen, "Deep exemplar-based video colorization," Proc. of IEEE Computer Vision and Pattern Recognition, pp.8044–8053, 2019.

[16] S. Iizuka and E. Simo-Serra, "DeepRemaster: Temporal source-reference attention networks for comprehensive video enhancement," ACM Trans. Graph. (SIGGRAPH), vol.38, no.6, p.176, 2019.

[17] Y.-Z. Zhao, L.-M. Po, W.-Y. Yu, Y.A.U. Rehman, M.-Y. Liu, Y.-J. Zhang, and W.-F. Ou, "VCGAN: Video colorization with hybrid generative adversarial network," IEEE Transactions on Multimedia, vol.25, pp.3017–3032, 2023.

[18] H. Liu, M. Xie, J. Xing, C. Li, C.-S. Leung, and T.-T. Wong, "ColorDiffuser: Video colorization with pretrained text-to-image diffusion models," Proc. of ACM International Conference Multimedia, pp.8891–8900, 2025.

[19] Z. Chang, S. Weng, H. Ouyang, L. Lin, Y. Li, S. Li, and B. Shi, "L-C4: Language-based video colorization for creative and consistent color," Neurocomputing, vol.665, p.132199, 2026. https://www.sciencedirect.com/science/article/pii/S0925231225028711

[20] Z. Fang, Z. Liu, K. Zhu, W. Zhai, Y. Cao, Y. Liu, K.L. Cheng, and Z.-J. Zha, "VanGogh: A unified multimodal diffusion-based framework for video colorization," 2025.