CONFIDENTIAL EXTENDED ABSTRACT.
DO NOT DISTRIBUTE ANYWHERE.

# 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[1,a)]　柳井 啓司[1,b)]

## 概要

本研究では，テキストプロンプトで色制御可能な長尺動画のカラー化手法を提案する．中央フレームから生成した参照画像を用いてアンカーフレームをカラー化し，その間を動画拡散モデルで補間する．さらに Lab 色空間で入力動画の輝度を復元することで，高画質かつ時間的に一貫したカラー動画を生成する．実験では既存手法を上回る知覚品質を示し，また長尺動画も一定のメモリ消費で処理可能であることを確認した．

## 1. はじめに

歴史映像や記録映像にはモノクロのまま保存されているものが多く，それらを自然にカラー化して保存・活用したいという需要は大きい．近年は深層学習により静止画カラー化が大きく進展し，拡散モデルの導入によって自然な発色やテキストによる色制御も可能になった．

一方，動画カラー化では時間的一貫性の確保が依然として難しい．各フレームを独立に生成するとフリッカーや色ドリフトが発生しやすく，さらに長尺動画ではフレーム数に応じて計算コストやメモリ消費が増大する．このため，高品質な色表現，時間的安定性，長尺動画への適用可能性を同時に満たす手法は依然として十分ではない．

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，長尺動画における一貫したカラー化を両立するフレームワークを提案する．提案手法では，中央フレームから生成した参照画像を用いてアンカーフレームを高品質にカラー化し，その間の中間フレームを動画拡散モデルにより補間することで，知覚品質と時間的一貫性を両立する．

## 2. 関連研究

### 2.1 動画カラー化

動画カラー化手法は，オプティカルフローに基づく色伝播を用いた手法 [2], [11]，ユーザー入力を用いるスクリブルベース [6], [12], [19]，参照画像の色を転写する参照ベース [4], [7], [21], [22]，エンドツーエンドの全自動手法 [10], [15], [24]，そして自然言語によるプロンプトベース手法 [3], [5], [13], [14] に大別される．従来手法では時間的一貫性を確保しやすい一方で発色表現が限定される場合があり，近年の生成モデルベース手法では高品質な発色が得られる反面，フレーム独立処理に起因する色揺れやドリフトが課題となる．

### 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り動画全体の色調を制御するプロンプトベース手法に位置づけられる．ただし，各フレームを直接テキストから生成するのではなく，テキストから得た中央フレームの参照画像を介して色を伝播させるハイブリッドな構成を採る点に特徴がある．基盤技術として，参照画像の生成には，Qwen-Image-Edit[20] を採用する．疎にサンプリングしたアンカーフレームでの色情報の転写には，Kosugi[8] が提案した Dual Attention 機構を導入した Stable Diffusion [18] を使用する．さらに，アンカーフレーム間の中間フレームの補間には Stable Video Diffusion (SVD) [1] を利用する．ここでは Liu ら [16] の枠組みに基づき，双方向推論と，マルチモーダル ControlNet による構造制御を組み合わせる．

## 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

提案手法の全体像を図 1 に示す．本手法は，参照画像生成，アンカーフレームのカラー化，中間フレーム補間，Lab 色空間での統合の四段階から構成される．

まず，入力動画の中央フレームを Qwen-Image-Edit でカラー化し，動画全体の色調の基準となる参照画像を生成する．次に，動画から一定間隔で抽出したアンカーフレームを，参照画像を用いて Dual Attention 付き Stable

[1] 電気通信大学
[a)] kimura-r@mm.inf.uec.ac.jp
[b)] yanai@cs.uec.ac.jp

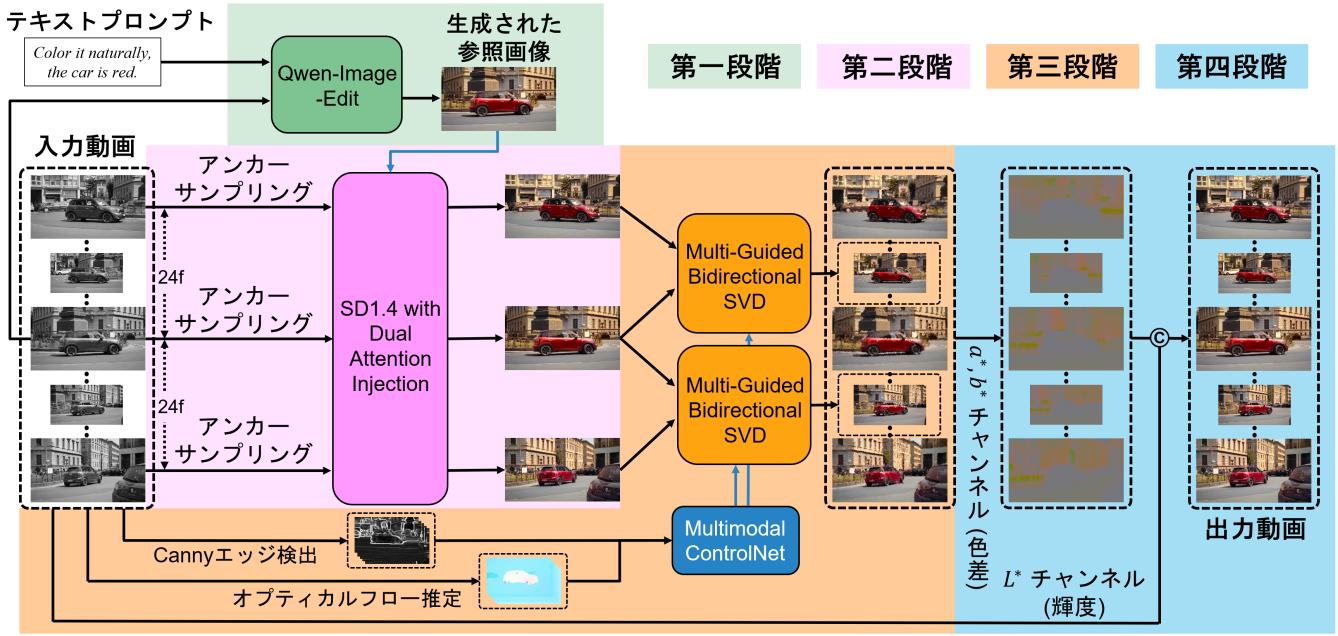



**図 1**　提案手法の全体パイプライン.

Diffusion でカラー化する．続いて，隣接するアンカーフレーム間の中間フレームを Stable Video Diffusion で補間し，その際に元のグレースケール動画から抽出したエッジとオプティカルフローを ControlNet で注入することで，構造と動きを拘束する．最後に，生成動画の色差成分と入力動画の輝度成分を Lab 色空間で統合し，元動画のディテールを保持した最終カラー動画を得る．

### 3.2　参照画像生成

動画全体で一貫した配色を与えるため，入力グレースケール動画から中央フレーム $v_{\text{mid}}$ を抽出し，これを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[20] を用い，$v_{\text{mid}}$ とプロンプト $T$ を入力して，構造を保ったカラー参照画像 $I_{\text{ref}}$ を生成する．得られた $I_{\text{ref}}$ 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，プロンプトの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

各アンカーフレームのカラー化には，Stable Diffusion 1.4 をベースとし，小杉ら [8] の Dual Attention 機構を導入する．入力フレーム，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を VAE で潜在表現へ変換し，DDIM Inversion により拡散過程を遡って各時刻の特徴を抽出する．

生成時には，入力と参照の構造対応を表す Gray-to-Gray Attention と，生成途中の特徴と参照色を結び付ける Colorized-to-Color Attention を統合する：

$$A_{\text{dual}} = \text{softmax}(S_{(\text{g2g})}\gamma + S_{(\text{c2c})}(1-\gamma)). \qquad (1)$$

ここで，$S_{(\text{g2g})}$ は入力フレームとグレースケール参照画像の対応を，$S_{(\text{c2c})}$ は生成途中の特徴と参照画像の色特徴の対応を表す．この Dual Attention により，入力構造を保持しながら，意味的に対応する領域へ参照画像の色を注入できる．

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．

また，さらに品質を向上させるため，必要に応じてこの生成後の画像を新たな初期値として扱い，再度エンコードして反転・生成を行う反復的洗練を実施する．この工程を経ることで，SD1.4 の高い生成能力を活かしつつ，参照画像と整合性の取れた高品質なアンカーフレームを得ることができる．

### 3.4　中間フレーム補間

前段階で得られるカラー化アンカーフレームは時間的に疎であるため，その間のフレームを補間し，時間方向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，Stable Video Diffusion (SVD)[1] を用いた条件付き動画生成を行う．ただし，単純な一方向生成では，開始フレームから離れるにつれて誤差が累積し，終点側のアンカーフレームとの不整合が生じやすい．この問題は長尺動画ほど顕著であり，時間方向に色調が徐々に崩れるドリフトの一因となる．

そこで本研究では，AnchorSync[16] の考え方に基づき，

表 1　DAVIS30 における定量的評価結果.

| Method | PSNR ↑ | SSIM ↑ | LPIPS ↓ | FID ↓ | CDC ↓ | Colorfulness ↑ |
|---|---|---|---|---|---|---|
| AutoColor (CVPR2019) [10] | 24.41 | 0.915 | 0.264 | 83.05 | 0.003734 | 14.14 |
| DeepExemplar (CVPR2019) [22] | 21.78 | 0.846 | 0.325 | 77.26 | 0.004006 | 28.82 |
| DeepRemaster (TOG2019) [7] | 21.95 | 0.848 | 0.354 | 97.54 | 0.005098 | 25.66 |
| TCVC (CVM2024) [15] | 25.17 | 0.921 | 0.239 | 74.94 | 0.003649 | 21.72 |
| VCGAN (TMM2023) [24] | 23.90 | 0.910 | 0.247 | 70.29 | 0.005303 | 15.89 |
| ColorDiffuser (ACMMM2025) [14] | 23.73 | 0.939 | 0.213 | 69.51 | 0.003607 | 29.13 |
| VanGogh (arXiv2025) [5] | 23.20 | 0.939 | 0.191 | – | – | **60.09** |
| L-C4 (arXiv2026) [3] | 25.69 | 0.933 | 0.209 | – | **0.003114** | 29.33 |
| **Ours** | **26.28** | **0.942** | **0.117** | **61.50** | 0.003965 | 43.51 |

表 2　Videvo における定量的評価結果.

| Method | PSNR ↑ | SSIM ↑ | LPIPS ↓ | FID ↓ | CDC ↓ | Colorfulness ↑ |
|---|---|---|---|---|---|---|
| AutoColor (CVPR2019) [10] | 25.90 | 0.925 | 0.277 | 76.28 | 0.001668 | 13.23 |
| DeepExemplar (CVPR2019) [22] | 20.63 | 0.831 | 0.348 | 76.63 | 0.002011 | 32.44 |
| DeepRemaster (TOG2019) [7] | 21.88 | 0.856 | 0.358 | 86.23 | 0.003607 | 28.72 |
| TCVC (CVM2024) [15] | 25.18 | 0.929 | 0.273 | 76.02 | 0.001629 | 18.89 |
| VCGAN (TMM2023) [24] | 24.67 | 0.919 | 0.276 | **63.83** | 0.002998 | 14.90 |
| ColorDiffuser (ACMMM2025) [14] | 25.27 | **0.951** | 0.205 | 66.11 | 0.001591 | 20.73 |
| L-C4 (arXiv2026) [3] | 25.17 | 0.939 | 0.198 | – | **0.001572** | 32.59 |
| **Ours** | **27.38** | 0.949 | **0.107** | 65.53 | 0.002446 | **38.98** |

隣接する 2 枚のアンカーフレームを始点・終点条件として用いる双方向補間を採用する．さらに，元のグレースケール動画から抽出した構造情報と動き情報を ControlNet[23] により SVD へ注入し，単なる見た目の補間ではなく，入力動画の形状変化や運動に沿ったカラー補間を行う．

#### 3.4.1　構造・動作制約の導入

単純な補間では，元動画の中間フレームと異なる形状や運動を持つ画像が生成される可能性がある．そこで本手法では，各フレームに対して Canny エッジとオプティカルフローを抽出し，これらをマルチモーダル ControlNet の条件入力として用いる．エッジは物体境界や輪郭の保持に寄与し，オプティカルフローは時間方向の動きの整合性を与える．

抽出された条件特徴は，SVD の U-Net 各ブロックへ注入される．これにより，生成される中間フレームは，アンカーフレームから受け継いだ色彩を保ちながらも，元のグレースケール動画に含まれる構造変化や運動パターンに忠実なものとなる．特に，人物や物体の移動が大きい場面でも，境界の破綻や不自然な色のはみ出しを抑制できる．

#### 3.4.2　双方向時間融合

一般的な SVD は，開始フレームのみを条件として時間順方向に生成を進めるため，シーケンス後半で誤差が累積しやすい．本手法ではこれを避けるために，始点アンカー $f'_i$ を条件とした順方向生成と，終点アンカー $f'_{i+1}$ を条件とした逆方向生成を並行して行う．順方向生成で得られる潜在表現を $Z^{\text{forward}}$，逆方向生成で得られる潜在表現を $Z^{\text{reverse}}$ とする．

各デノイジング段階において，両者の潜在表現を以下のように融合する：

$$Z_j^{\text{fuse}} = \alpha Z_j^{\text{forward}} + (1-\alpha) Z_j^{\text{reverse}}. \tag{2}$$

ここで $\alpha$ は時間位置に応じた重みであり，始点付近では順方向の寄与を大きく，終点付近では逆方向の寄与を大きくする．この融合により，中間フレーム列は始点・終点の両アンカーフレームと整合し，区間全体として滑らかな遷移を実現する．

さらに，この双方向融合は色の時間的安定化にも寄与する．一方向生成では後半に向かうほど色が徐々に変化する傾向があるのに対し，本手法では終点側のアンカーフレームからも制約が与えられるため，区間全体で色調のブレを抑えられる．この処理をすべてのアンカー区間に対して繰り返し，区間ごとの結果を連結することで，長い動画全体に対しても安定したカラー化を実現する．

### 3.5　Lab 色空間での統合

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高周波成分や細部構造を保持しており，構造情報としての信頼性が高い．

そこで本手法では，生成動画 $V_{\text{gen}}$ と入力動画 $V_{\text{in}}$ を Lab 色空間に変換し，$V_{\text{gen}}$ から色差成分 $a^*, b^*$ を，$V_{\text{in}}$ から輝度成分 $L^*$ を抽出して再結合する．これにより，生成された色彩を維持しつつ，元動画の解像感と細部構造を復元した最終動画を得る．

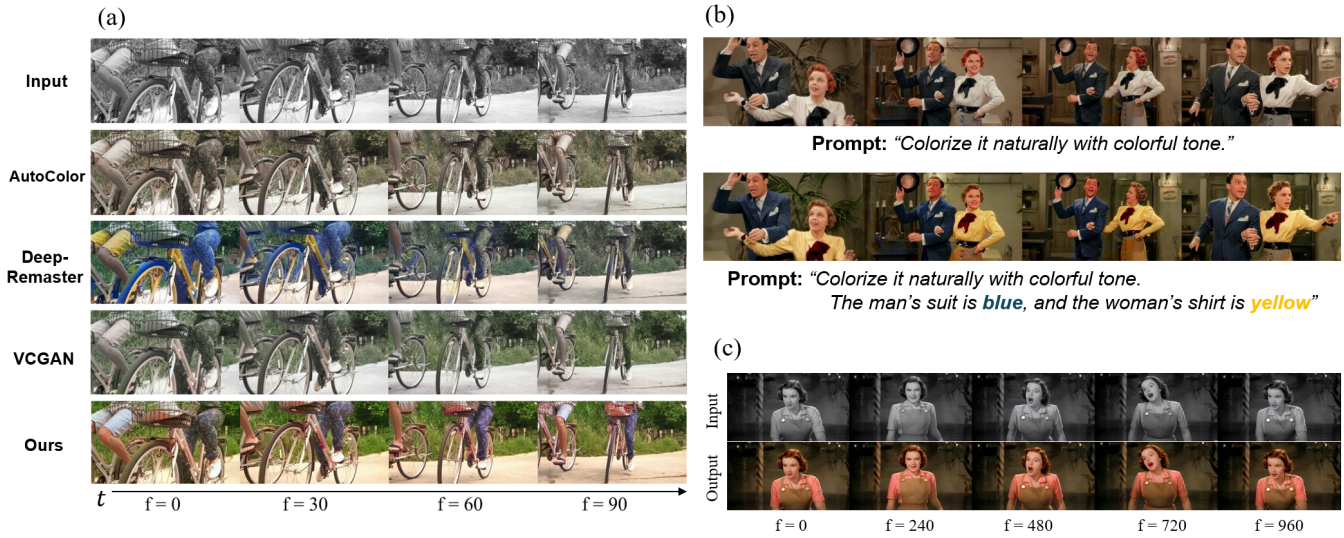



**図 2** (a) 従来手法と提案手法によるカラー化結果の比較．(b) 異なるプロンプトによる生成結果の比較．(c) 長尺動画のカラー化結果．

## 4. 実験

### 4.1 実験設定

評価には DAVIS30[17] および Videvo[9] を用いた．アンカー間隔は $K = 24$ とし，共通プロンプトとして *"Colorize it naturally with colorful tone."* を用いた．評価指標には PSNR，SSIM，LPIPS，FID，CDC，Colorfulness を採用し，画質・知覚品質・時間的一貫性を多面的に評価した．比較対象には AutoColor，DeepExemplar，DeepRemaster，TCVC，VCGAN，ColorDiffuser，L-C4，VanGogh を用いた．

### 4.2 定量的評価

DAVIS30 および Videvo における定量評価では，提案手法は PSNR，SSIM，LPIPS，FID の主要指標で高水準の性能を示した．特に LPIPS は DAVIS30 で 0.117，Videvo で 0.107 と最良であり，知覚品質の高さが確認できた．また，Colorfulness も高く，鮮やかさと自然さを両立している．一方，CDC では時間的一貫性に特化した手法に及ばない場合もあるが，双方向補間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た．

### 4.3 定性的評価

図 2(a) に従来手法との比較結果を示す．AutoColor[10] や VCGAN[24] では全体的に彩度が低く色が滲む傾向があり，DeepRemaster[7] では時間の経過とともに色の一貫性が失われる．一方，提案手法は参照画像の色情報を安定して反映し，アンカーフレーム間でも色調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 による構造拘束が有効に働き，動きのある場面でもエッジの破綻が少なく，滑らかなカラー化が得られた．

#### 4.3.1 プロンプトによる制御

図 2(b) 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．提案手法では，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用いているため，プロンプトを変更するだけで，ユーザーの意図に応じた配色を動画全体へ一貫して適用できることが確認された．

#### 4.3.2 長尺動画への適用

図 2(c) は，300 フレームを超える長尺動画を用いた結果である．従来の動画カラー化では，動画長に比例して計算コストやメモリ消費が増大することが課題であった．これに対し提案手法では，動画全体をアンカー区間ごとに分割して処理するため，メモリ消費量を約 22GB に抑えつつ長尺動画を処理できる．結果として，長いシーケンス全体を通しても色調のブレや破綻が少なく，安定したカラー化が実現された．

## 5. おわりに

本研究では，テキストプロンプトによる色制御と時間的一貫性を両立する長尺動画カラー化手法を提案した．中央フレームから生成した参照画像を用いた Dual Attention ベースのアンカーフレームカラー化と，ControlNet 付き SVD による双方向補間を組み合わせることで，高い知覚品質と時間的安定性を実現した．DAVIS30 および Videvo での実験により有効性を確認し，長尺動画も一定のメモリ消費量で処理可能であることを示した．今後は，複数参照フレームの利用，激しい運動や多物体シーンへの対応，シーンチェンジを含む動画への拡張を進める．

## 参考文献


[1] Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., Jampani, V. and Rombach, R.: Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets (2023). arXiv:2311.15127.

[2] Bonneel, N., Tompkin, J., Sunkavalli, K., Sun, D., Paris, S. and Pfister, H.: Blind Video Temporal Consistency, *ACM Trans. Graph. (SIGGRAPH)*, Vol. 34, No. 6, pp. 196–204 (2015).

[3] Chang, Z., Weng, S., Ouyang, H., Li, Y., Li, S. and Shi, B.: L-C4: Language-based video colorization for creative and consistent color, *Neurocomputing*, Vol. 665, p. 132199 (2026).

[4] Fang, X., Dai, L. and Tang, J.: OmniFusion: Exemplar-based Video Colorization Using OmniMotion and Diffusion Priors, *Proc. of Asian Conference on Computer Vision*, pp. 1215–1232 (2024).

[5] Fang, Z., Liu, Z., Zhu, K., Zhai, W., Cao, Y., Liu, Y., Cheng, K. L. and Zha, Z.-J.: VanGogh: A Unified Multimodal Diffusion-based Framework for Video Colorization (2025). arXiv:2501.09499.

[6] Heu, J.-H., Hyun, D.-Y., Kim, C.-S. and Lee, S.-U.: Image and Video Colorization Based on Prioritized Source Propagation, *Proc. of IEEE International Conference on Image Processing*, pp. 465–468 (2009).

[7] Iizuka, S. and Simo-Serra, E.: DeepRemaster: Temporal Source-Reference Attention Networks for Comprehensive Video Enhancement, *ACM Trans. Graph. (SIGGRAPH)*, Vol. 38, No. 6, p. 176 (2019).

[8] Kosugi, S.: 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).

[9] Lai, W.-S., Huang, J.-B., Wang, O., Shechtman, E., Yumer, E. and Yang, M.-H.: Learning Blind Video Temporal Consistency, *Proc. of European Conference on Computer Vision*, Cham, pp. 179–195 (2018).

[10] Lei, C. and Chen, Q.: Fully Automatic Video Colorization with Self-Regularization and Diversity, *Proc. of IEEE Computer Vision and Pattern Recognition*, pp. 3748–3756 (2019).

[11] Lei, C., Xing, Y., Ouyang, H. and Chen, Q.: Deep Video Prior for Video Consistency and Propagation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 45, No. 1, pp. 356–371 (2023).

[12] Levin, A., Lischinski, D. and Weiss, Y.: Colorization Using Optimization, *ACM Trans. Graph. (SIGGRAPH)*, Vol. 23, No. 3, pp. 689–694 (2004).

[13] Li, J., Zhao, H., Wang, Y. and Lin, J.: Towards Photorealistic Video Colorization via Gated Color-Guided Image Diffusion Models, *Proc. of ACM International Conference on Multimedia*, pp. 10891–10900 (2024).

[14] Liu, H., Xie, M., Xing, J., Li, C., Leung, C.-S. and Wong, T.-T.: ColorDiffuser: Video Colorization with Pre-trained Text-to-Image Diffusion Models, *Proc. of ACM International Conference on Multimedia*, pp. 8891–8900 (2025).

[15] Liu, Y.-H., Zhao, H.-Y., Chan, K. C., Wang, X.-T., Loy, C. C., Qiao, Y. and Dong, C.: Temporally Consistent Video Colorization with Deep Feature Propagation and Self-Regularization Learning, *Computational Visual Media*, Vol. 10, No. 2, pp. 375–395 (2024).

[16] Liu, Z., Wang, Y., Wei, T. and Ma, C.: AnchorSync: Global Consistency Optimization for Long Video Editing, *Proc. of ACM International Conference on Multimedia*, pp. 4494–4503 (2025).

[17] Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M. and Sorkine-Hornung, A.: A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation, *Proc. of IEEE Computer Vision and Pattern Recognition*, pp. 724–732 (2016).

[18] Rombach, R., Blattmann, A., Lorenz, D., Esser, P. and Ommer, B.: High-Resolution Image Synthesis with Latent Diffusion Models, *Proc. of IEEE Computer Vision and Pattern Recognition*, pp. 10684–10695 (2022).

[19] Sheng, B., Sun, H., Magnor, M. and Li, P.: Video Colorization Using Parallel Optimization in Feature Space, *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 24, No. 3, pp. 407–417 (2014).

[20] Wu, C., Li, J. et al.: Qwen-Image Technical Report (2025). arXiv:2508.02324.

[21] Yang, Y., Pan, J., Peng, Z., Du, X., Tao, Z. and Tang, J.: BiSTNet: Semantic Image Prior Guided Bidirectional Temporal Feature Fusion for Deep Exemplar-Based Video Colorization, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 46, No. 8, pp. 5612–5624 (2024).

[22] Zhang, B., He, M., Liao, J., Sander, P. V., Yuan, L., Bermak, A. and Chen, D.: Deep Exemplar-based Video Colorization, *Proc. of IEEE Computer Vision and Pattern Recognition*, pp. 8044–8053 (2019).

[23] Zhang, L. and Agrawala, M.: Adding Conditional Control to Text-to-Image Diffusion Models, *Proc. of IEEE International Conference on Computer Vision*, pp. 3813–3824 (2023).

[24] Zhao, Y.-Z., Po, L.-M., Yu, W.-Y., Rehman, Y. A. U., Liu, M.-Y., Zhang, Y.-J. and Ou, W.-F.: VCGAN: Video Colorization with Hybrid Generative Adversarial Network, *IEEE Transactions on Multimedia*, Vol. 25, pp. 3017–3032 (2023).