拡散モデルを用い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

はじめに

はじめに – 背景

- 歴史的資料や記録映像の多くはモノクロのまま残されている。
- カラー化の需要は高い。

カラー化の対象

- **静止画**
  時間方向の制約がない。
- **動画**
  時間方向の制約がある。

f = 0　　f = 20　　f = 40

はじめに – 背景

動画のカラー化

■ コンピュータ技術の発展とともに応用が広がってきた。

■ 近年では生成モデルを活用した新たな手法が提案されている。

■ **共通の課題**
- 自然な発色
- 時間的一貫性
- 色指定の柔軟性

■ **最新手法の課題**
- 長尺動画での計算量増大

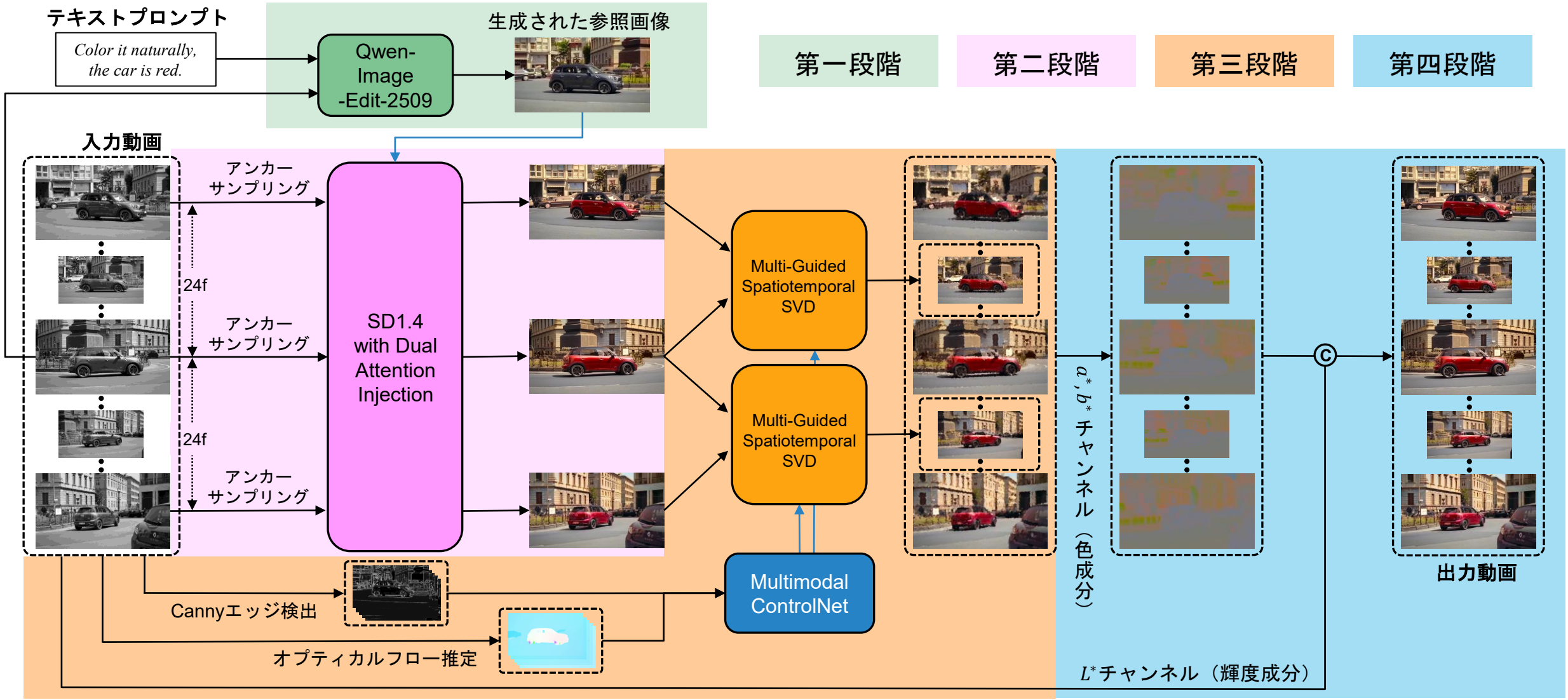
はじめに – 目的

目的

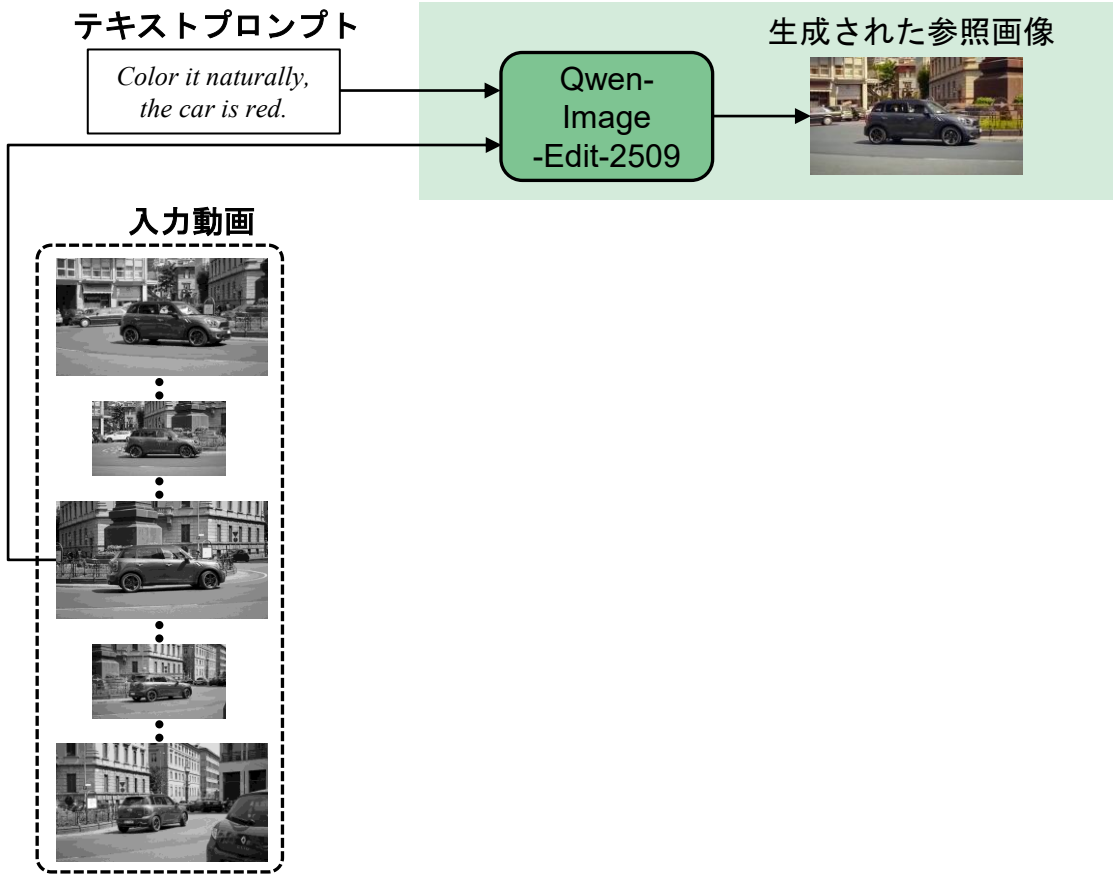
- 本研究では、数百フレームを超える長尺動画のカラー化を目的とする。
- 大規模に事前学習された画像・動画生成モデルを活用。
- 以下を満たす新たなフレームワークを提案。
  - 高い知覚品質
  - 時間的安定性
  - テキストプロンプトによる色制御

提案手法

提案手法 – 概要

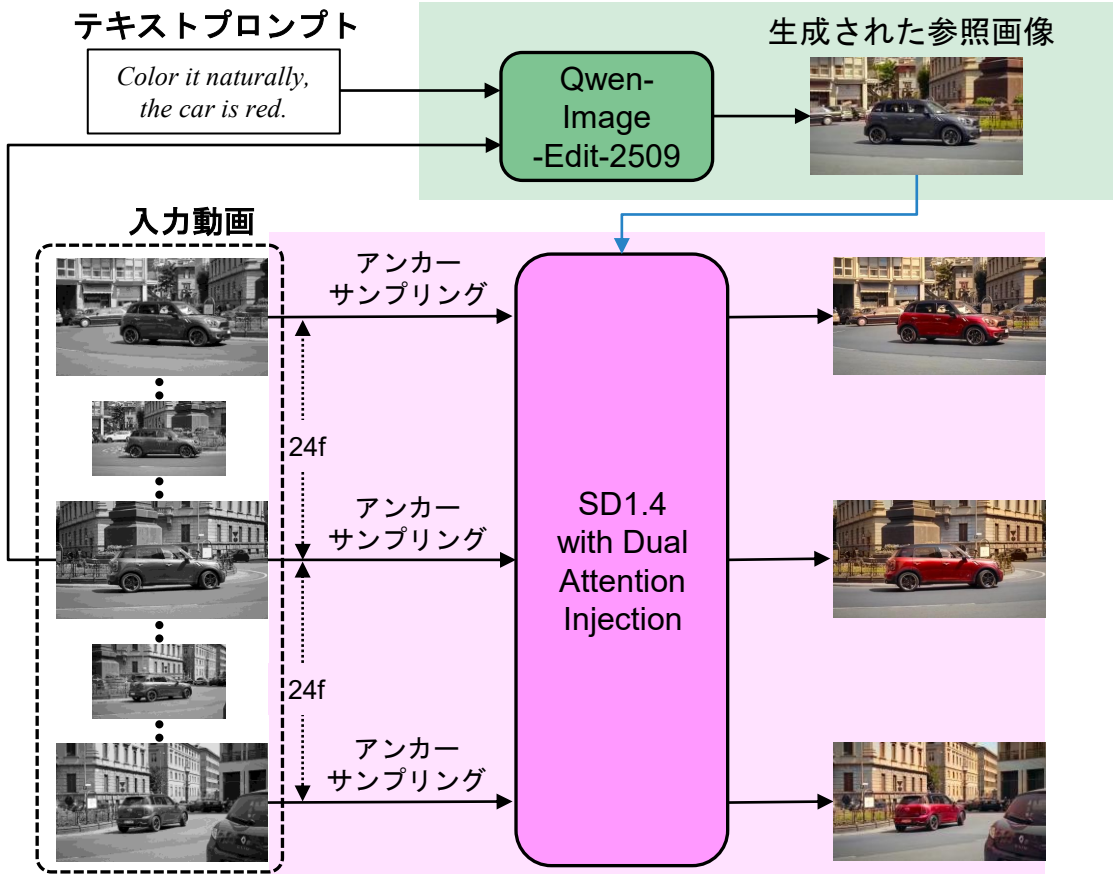

テキストプロンプト
Color it naturally, the car is red.
Qwen-Image-Edit-2509
生成された参照画像
第一段階
第二段階
第三段階
第四段階
入力動画
アンカーサンプリング
24f
アンカーサンプリング
24f
アンカーサンプリング
SD1.4 with Dual Attention Injection
Multi-Guided Spatiotemporal SVD
Multi-Guided Spatiotemporal SVD
Multimodal ControlNet
Cannyエッジ検出
オプティカルフロー推定
a^*, b^* チャンネル（色成分）
L^* チャンネル（輝度成分）
出力動画

提案手法 – 概要


テキストプロンプト
Color it naturally, the car is red.
Qwen-Image-Edit-2509
生成された参照画像
入力動画

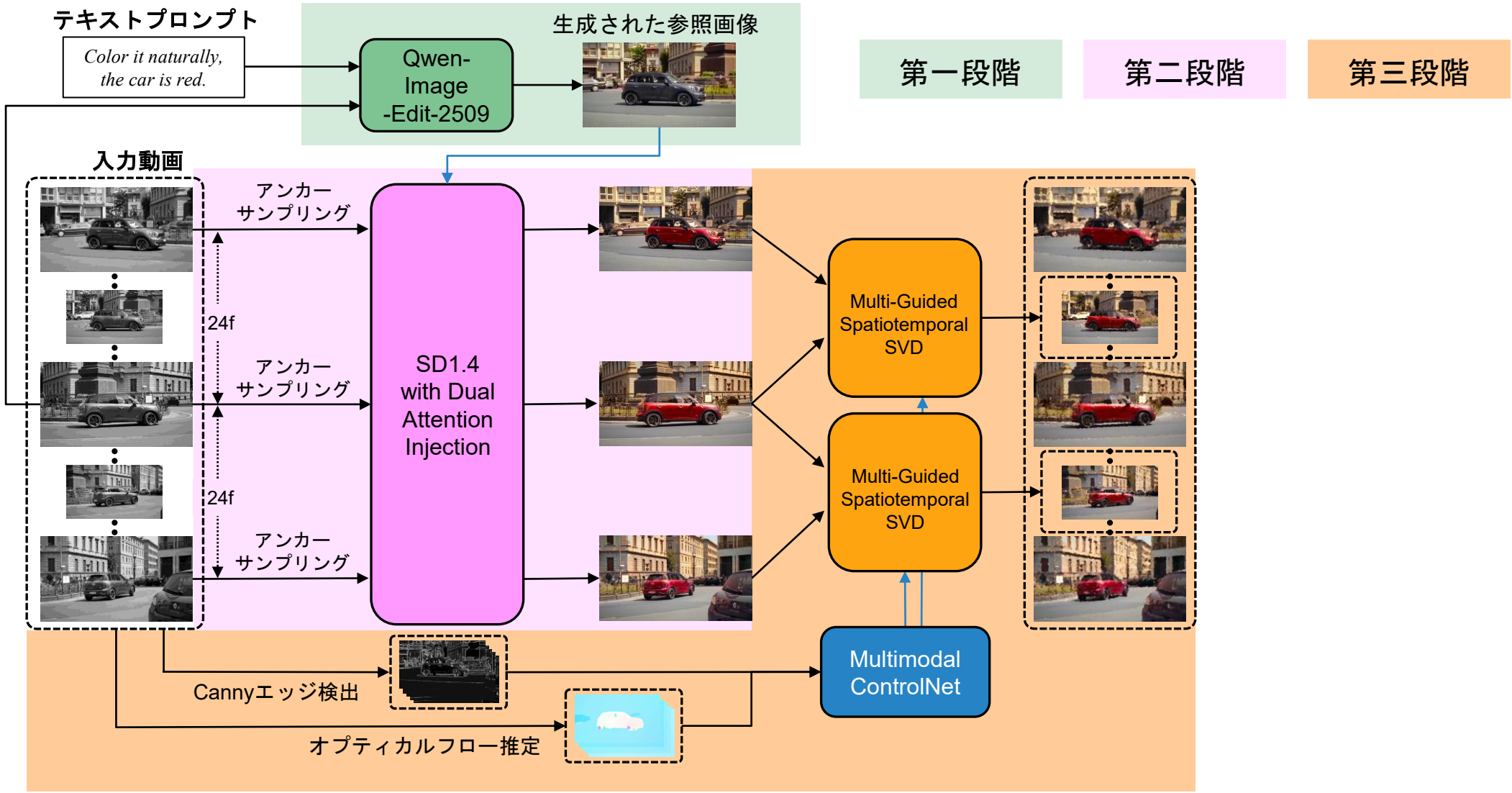
第一段階

提案手法 – 概要

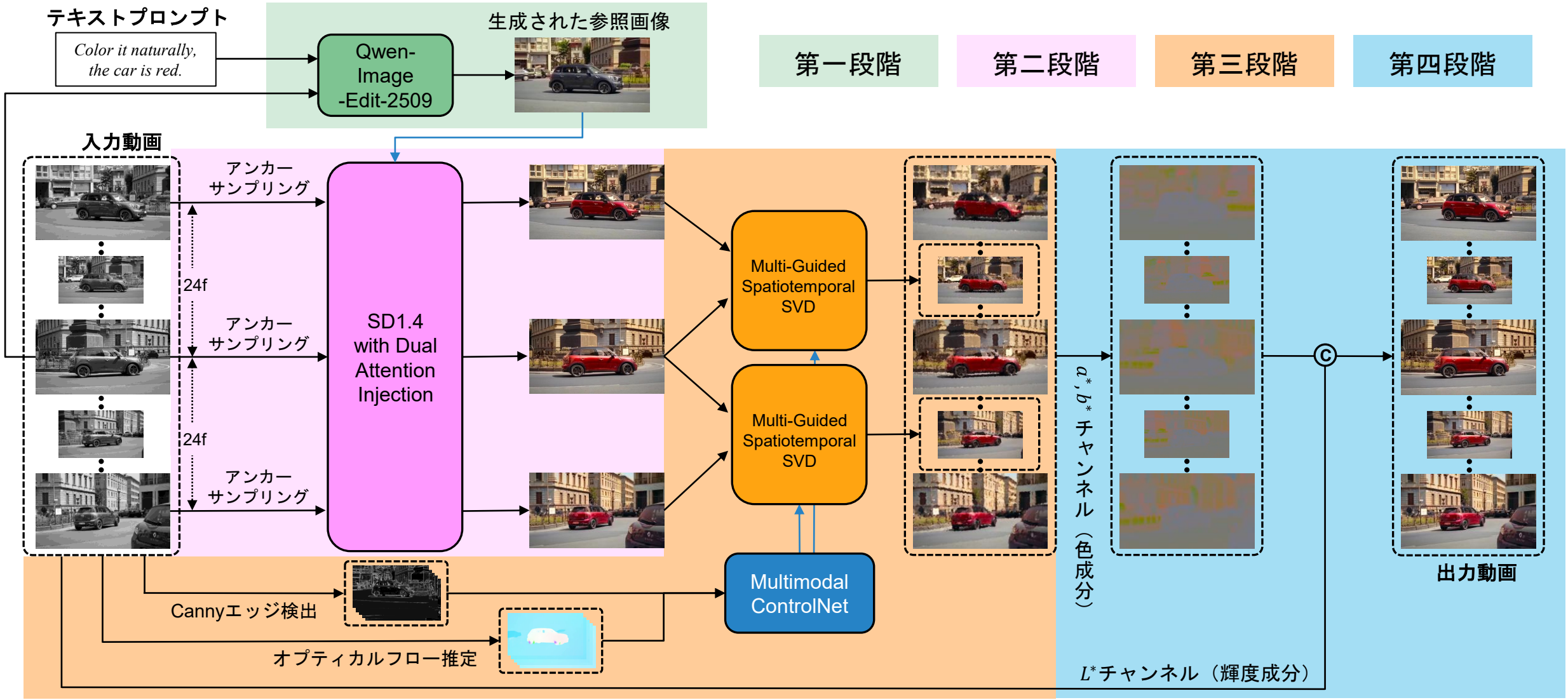


第一段階　第二段階

提案手法 – 概要


テキストプロンプト
Color it naturally, the car is red.
Qwen-Image-Edit-2509
生成された参照画像
第一段階
第二段階
第三段階
入力動画
アンカーサンプリング
24f
アンカーサンプリング
24f
アンカーサンプリング
SD1.4 with Dual Attention Injection
Multi-Guided Spatiotemporal SVD
Multi-Guided Spatiotemporal SVD
Multimodal ControlNet
Cannyエッジ検出
オプティカルフロー推定

提案手法 – 概要


テキストプロンプト
Color it naturally, the car is red.
Qwen-Image-Edit-2509
生成された参照画像
第一段階
第二段階
第三段階
第四段階
入力動画
アンカーサンプリング
24f
アンカーサンプリング
24f
アンカーサンプリング
SD1.4 with Dual Attention Injection
Multi-Guided Spatiotemporal SVD
Multi-Guided Spatiotemporal SVD
Multimodal ControlNet
Cannyエッジ検出
オプティカルフロー推定
a^*, b^*チャンネル（色成分）
L^*チャンネル（輝度成分）
出力動画

提案手法 – 第一段階（参照画像生成）

第一段階の目的

■ 高品質な参照画像を生成すること。

➡Qwen-Image-Editを採用。

テキストプロンプト

Color it naturally with colorful tone.

入力動画の中央フレーム

Qwen-Image -Edit-2509

カラー化された中央フレーム

➡参照画像

提案手法 – 第二段階（アンカーフレームのカラー化）

第二段階の目的

■ アンカーフレームに対し、参照画像の色情報を転写すること。

ベースモデル

■ Stable Diffusion 1.4 (SD1.4) を使用。

提案手法 – 第二段階（アンカーフレームのカラー化）

抽出されたアンカーフレーム

DDIM Inversion

z^{in}

Denoising U-Net

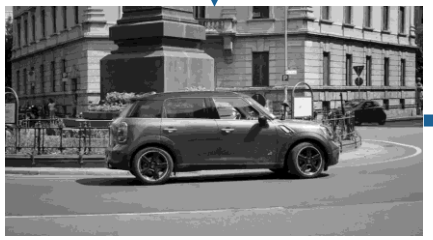
カラー化された
アンカーフレーム

生成された参照画像

DDIM Inversion

z^{ref}

L^*

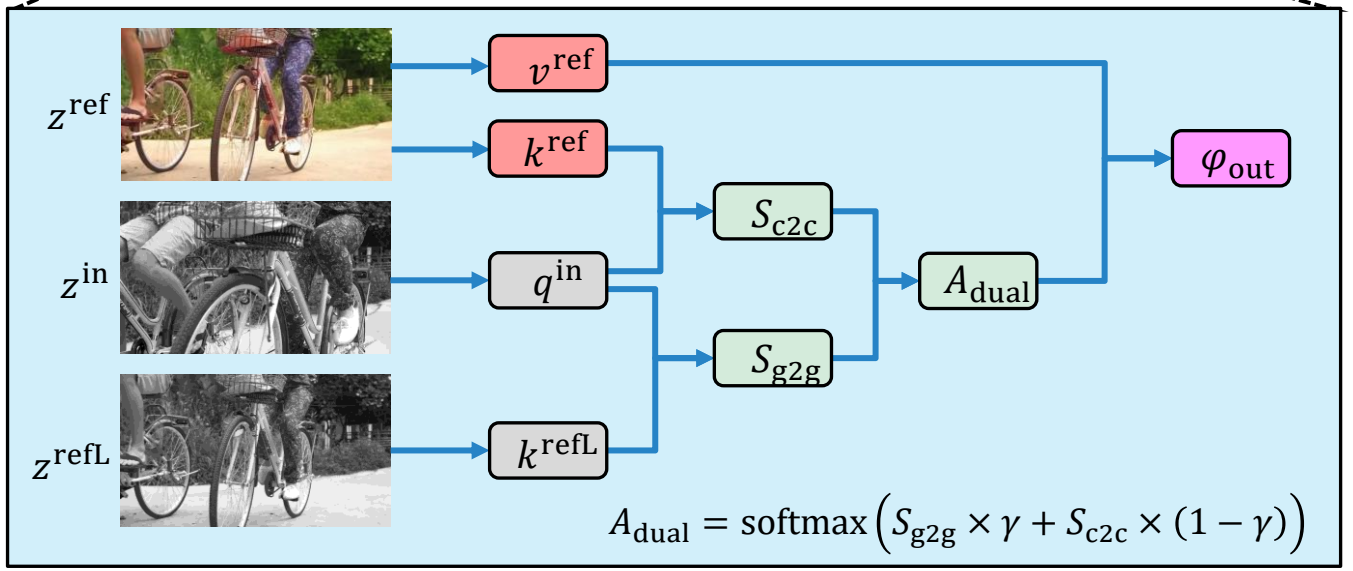


DDIM Inversion

z^{refL}

参照画像の輝度成分

Dual Attention機構



提案手法 – 第二段階（アンカーフレームのカラー化）

Classifier-free Colorization Guidance

- 条件付き推論と無条件推論の差分を増幅。

$$\widetilde{\epsilon_\theta}(z_t^{\text{out}}) = \epsilon_\theta^{\text{col}}(z_t^{\text{out}}) \times w + \epsilon_\theta(z_t^{\text{out}}) \times (1 - w)$$

反復的洗練と初期化

- 生成画像を再度エンコードして反転・生成を繰り返す。
- AdaIN（Adaptive Instance Normalization）を適用。

提案手法 – 第三段階（中間フレーム補間）

第三段階の目的

■ アンカーフレーム間の中間フレームを生成し、動画を再構成すること。

ベースモデル

■ Stable Video Diffusion (SVD)を使用。

- SVD単体ではドリフト現象が発生する。
- 元動画の構造や動きが反映されない恐れがある。

提案手法 – 第三段階（中間フレーム補間）

入力動画

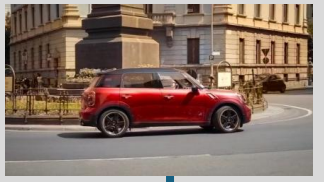
※全アンカーフレーム間で行う ➡

Canny
エッジ検出

オプティカル
フロー推定

Multimodal
ControlNet

カラー化された隣り合う2枚のアンカーフレーム



Multi-Guided Spatiotemporal
SVD

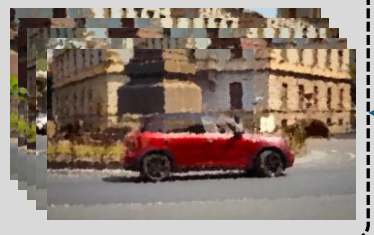
weightは共有

Forward Path

Backward Path

線形結合

補間された動画



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

提案手法 – 第四段階（LAB空間での統合）

第四段階の目的

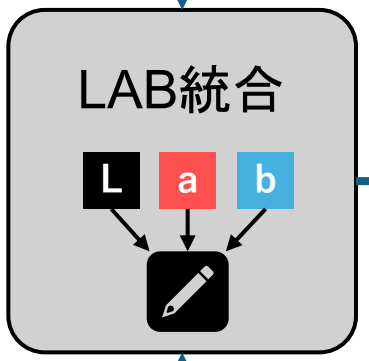
■ 生成過程で失われた微細な構造を復元すること。

提案手法 – 第四段階（LAB空間での統合）

生成動画（細かい構造がボケている）

元動画

a^*, b^*チャンネル



L^*チャンネル

最終出力動画

実験

実験 – 定量評価

実験設定

■ **データセット**

- DAVIS30（平均フレーム数: 57.8）
- Videvoデータセット（平均フレーム数: 226.6）

■ **評価指標**

- PSNR: 画素レベルでの信号忠実度を評価。
- LPIPS: 知覚的な類似度を評価。
- FID: 画像のリアリズムを評価。
- Colorfulness: 色彩の豊かさを定量化する指標。

実験 – 定量評価

実験結果

■ DAVIS30データセットでの評価結果。（**太字**が最良、<u>下線</u>が次点）

Method	色制御の方法	PSNR ↑	LPIPS↓	FID↓	ColorfulNess↑
AutoColor (2019)	全自動	24.41	0.264	83.05	14.14
DeepRemaster (2019)	参照画像	21.95	0.354	97.54	25.66
VCGAN (2021)	全自動	23.90	0.247	70.29	15.89
ColorDiffuser (2025)	プロンプト	23.73	0.213	<u>69.51</u>	29.13
L-C4 (2026)	プロンプト	<u>25.69</u>	0.209	データなし	29.33
VanGogh (2025)	プロンプト	23.20	<u>0.191</u>	データなし	**60.09**
提案手法	プロンプト	**26.28**	**0.117**	**61.50**	<u>43.51</u>

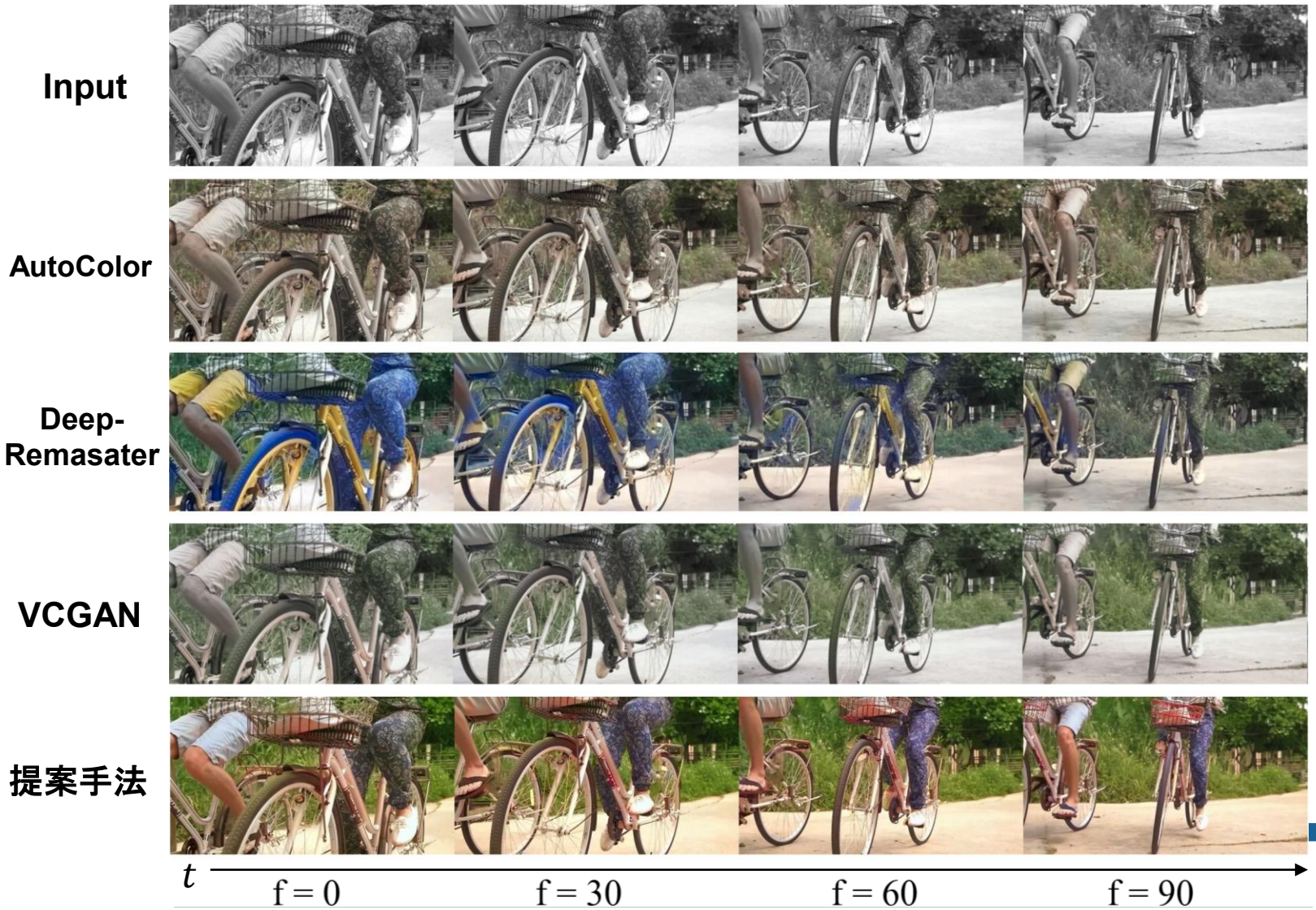
実験 – 定量評価

実験結果

■ Videvoデータセットでの評価結果。（**太字**が最良、<u>下線</u>が次点）

Method	色制御の方法	PSNR ↑	LPIPS↓	FID↓	ColorfulNess↑
AutoColor (2019)	全自動	<u>25.90</u>	0.277	76.28	13.23
DeepRemaster (2019)	参照画像	21.88	0.358	86.23	28.72
VCGAN (2021)	全自動	24.67	0.276	**63.83**	14.90
ColorDiffuser (2025)	プロンプト	25.27	0.205	66.11	20.73
L-C4 (2026)	プロンプト	25.17	<u>0.198</u>	データなし	<u>32.59</u>
VanGogh (2025)	プロンプト	データなし	データなし	データなし	データなし
提案手法	プロンプト	**27.38**	**0.107**	<u>65.53</u>	**38.98**

実験 – 定性評価



色制御の方法

全自動

参照画像

全自動

プロンプト

提案手法

実験 – 定性評価

プロンプトによる柔軟な色指定

■ 色を指定するプロンプトを与えることで柔軟なカラー化が可能。

Color it naturally with colorful tone.

*Color it naturally with colorful tone. The man's suit is **blue**, and the woman's shirts is **yellow**.*

Color it naturally with colorful tone.

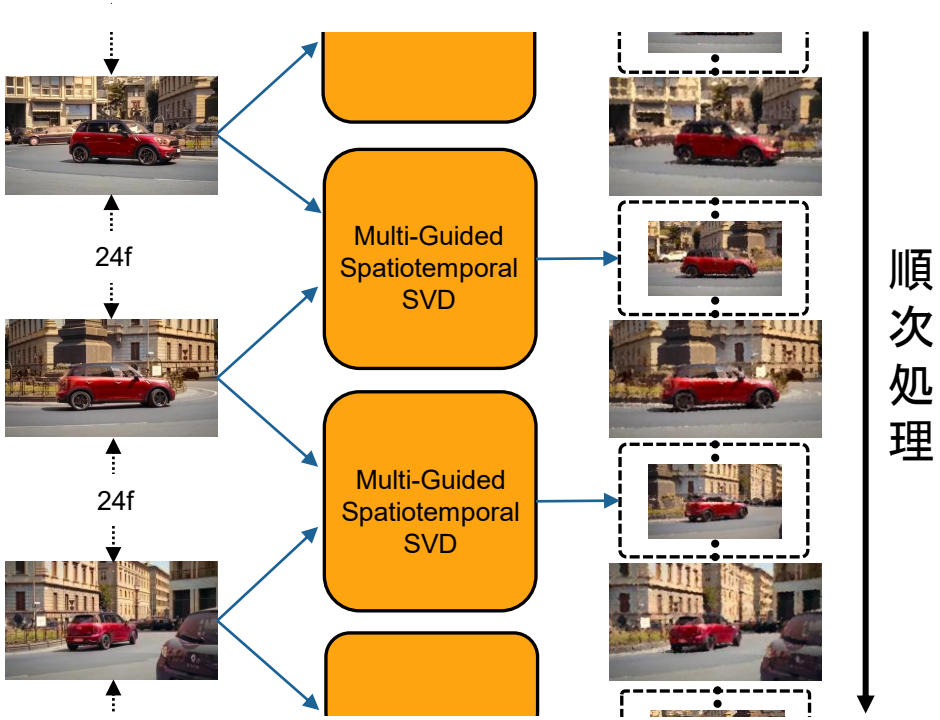
*Color it naturally with colorful tone. The cat has **purple** fur and **yellow** eyes.*

実験 – 定性評価

長尺動画への対応

■ 最新の従来手法では数百フレームを超える長尺動画を処理できない。

■ 提案手法はメモリ使用量を抑え、長尺動画でも安定して処理可能。



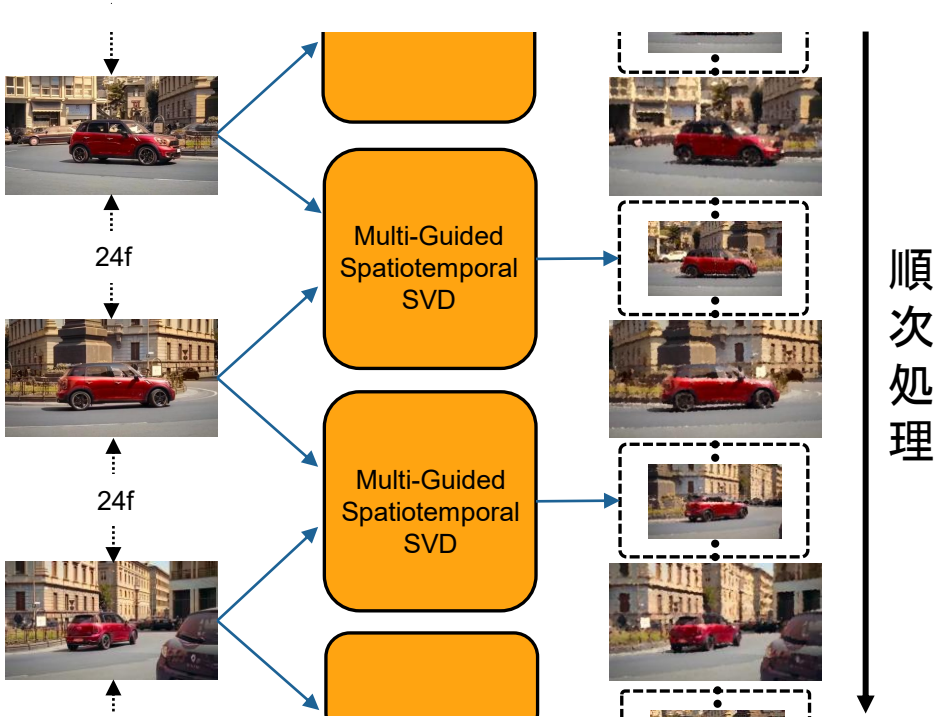
全960フレームの長尺動画

実験 –

長尺動画への対応

■ 最新の従来手法では数百フレームを超える長尺動画を処理できない。

■ 提案手法はメモリ使用量を抑え、長尺動画でも安定して処理可能。



全960フレームの長尺動画

アブレーションスタディ– 中間フレーム補完の効果

■ 中間フレーム補間を行わず、全フレームを独立して処理。

f = 0　　f = 1　　f = 3

アンカーフレーム間隔	実行時間※ (s)
K=1	724
K=24 (Ours)	476

※第二段階と第三段階の合計実行時間

アブレーションスタディ– 双方向推論の効果

入力動画

w/o ControlNet
第三段階の出力

w/o ControlNet
最終出力動画

w/ ControlNet
最終出力動画

f = 24（アンカーフレーム）　f = 32　f = 40　f = 48（アンカーフレーム）

アブレーションスタディー マルチモーダルControlNetの効果

入力動画

w/o 逆方向推論
第三段階の出力

w/o 逆方向推論
最終出力動画

w/ 逆方向推論
最終出力動画

f = 22　　f = 23　　f = 24（アンカーフレーム）　　f = 25

課題 – アスペクト比の制約

- 9:5以外のアスペクト比を入力すると補間が崩壊。

入力動画

第三段階の出力

最終出力動画

課題 – シーン切り替え

- シーン切り替えには未対応。
- シーン検知を使って対応させても一貫性が損失。

まとめ

まとめ

まとめ

■ 長尺動画を扱える新たな動画カラー化手法を提案。

■ 計算コストを抑えつつ、柔軟で高精度なカラー化に成功。

■ 従来手法を上回る性能を達成、長尺動画も処理可能であることを示した。

課題と今後の展望

■ 解像度の制約、遮蔽領域の問題、シーン切り替え時の不整合が課題。

■ 可変解像度への対応、動的に参照画像を更新する機構の導入、高速化。