Échantillonnage de groupes de trames pour une édition de vidéos zero-shot par modèle de diffusion

Thérèse TISSEAU DES ESCOTAIS[1, 2] Clément RAMBOUR[1, 3] Bertrand LEROY[2] Arnaud BRELOY[1]

[1]Centre d'Etudes et De Recherche en Informatique et Communications (CEDRIC), 2 rue Conté, 75003 Paris, France

[2]Ampere Software Technology, 122 - 122 122 B Avenue du Général Leclerc, 92100 Boulogne-Billancourt, France

[3]Institut des Systèmes Intelligents et Robotique (ISIR), Pyramide - T55, 4 Place Jussieu 65, 75005 Paris, France

Résumé – L'enjeu de l'édition de vidéos à partir d'un texte est de générer une vidéo qui respecte la vidéo originale et le texte d'entrée, tout en gardant une cohérence temporelle entre les trames. Dans ce contexte, nous proposons une méthode d'édition de vidéos zero-shot utilisant un modèle de diffusion d'édition d'images pré-entraîné. Pour induire une attention temporelle dans ce processus, les trames sont traitées par grilles construites via des permutations aléatoires : nous proposons ici d'utiliser une loi de Mallows pour assurer un bon compromis entre la cohérence globale de l'édition et la préservation des détails. Des expériences sur le jeu de données DAVIS [8] montrent l'intérêt de l'approche proposée.

Abstract – The main challenge in video editing is to generate a video that remains faithful to both the input video and the input text, while maintaining temporal consistency between frames. In this context, we propose a zero-shot video editing method that leverages a pre-trained text-to-image diffusion model. To introduce temporal attention in this process, frames are processed using grids constructed via random permutations. Here, we propose to use a Mallows distribution to achieve a good balance between overall editing coherence and detail preservation. Experiments on DAVIS [8] dataset demonstrate the efficiency of the proposed approach.

1 Introduction

Les modèles de diffusion ont récemment permis d'obtenir des résultats impressionnants dans le domaine de l'édition d'images par texte [10, 13]. En contraste, l'entraînement de modèles d'édition de vidéos reste un défi, car la dimension des données et leur nature spatio-temporelle posent des limitations en terme de ressources computationnelles.

Pour cette raison, une partie de la communauté se tourne vers le développement d'approches zero-shot [1, 9, 5, 2, 6, 4, 12] réutilisant des modèles d'édition d'images pré-entraînés. Dans cette perspective, l'objectif principal est de proposer des architectures assurant naturellement une cohérence temporelle dans l'édition, en vue notamment d'éviter des phénomènes de *flickering* (apparition et disparition d'objets dans la scène). Les méthodes récentes se basent principalement sur la mise en œuvre (plus ou moins implicite) d'une attention temporelle au travers des similarités entre les trames [2, 6, 12]. Une autre approche récente, nommée RAVE [4], obtient astucieusement cette attention temporelle via une attention spatiale calculée sur la concaténation de plusieurs trames dans une grille. Pour assurer un style d'édition constant sur toute la vidéo, les grilles ne sont pas formées par une concaténation fixée de trames consécutives, mais par des tirages aléatoires ré-échantillonnés tout au long du processus de diffusion. La méthode [4] parvient ainsi à produire des éditions parmi les plus satisfaisantes de l'état de l'art. Cependant, elle atteint ses limites quand les vidéos sont longues ou quand la scène présente trop de mouvement, car l'édition conjointe de trames trop différentes provoque généralement les phénomènes de *flickering* précédemment mentionnés.

Ce travail vise à pallier la limitation précédemment évoquée en introduisant un contrôle plus fin du tirage aléatoire déterminant la formation des grilles. Nous proposons de nous appuyer sur la loi de Mallows [7] pour raffiner ce tirage : cette approche offre un compromis paramétré entre une permutation aléatoire uniforme (équivalent à la méthode RAVE [4]) et une permutation se restreignant aux trames adjacentes. Une série d'expériences sur le jeu de données DAVIS [8] montre que l'approche proposée permet d'améliorer les métriques d'évaluation utilisées par l'état de l'art, ainsi que les résultats qualitatifs visuels.

2 Préliminaires

2.1 Edition d'images par prompt

Les méthodes d'édition de vidéos à la pointe utilisent le processus de diffusion [3, 11] sur les images. Le processus de diffusion appliqué à une image comprend deux parties : un processus direct qui prend en entrée une image et la bruite graduellement en un bruit aléatoire, et un processus retour qui apprend à retrouver l'image d'entrée à partir du bruit aléatoire généré. Ainsi, les méthodes d'édition d'images par diffusion apprennent à débruiter un bruit aléatoire en une image structurée, guidées par un *prompt* (instruction textuelle) d'entrée.

Pour l'édition de vidéos, l'entraînement et l'affinement de modèles sont très coûteux. C'est pourquoi beaucoup de méthodes zero-shot se développent. Les méthodes d'édition de vidéos zero-shot s'appuient sur des méthodes d'édition d'images pré-entraînées [10, 13] pour éditer toutes les images de la vidéo d'entrée.

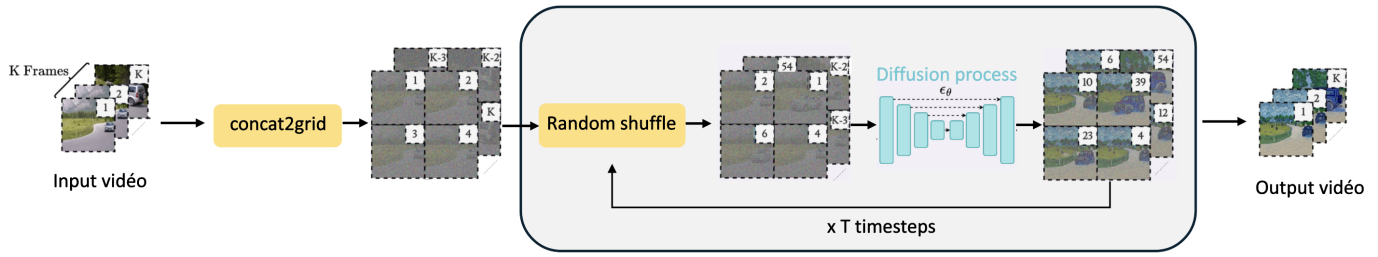

FIGURE 1 : **Architecture de notre méthode.** Deux étapes principales : concaténation des trames en grilles ; permutation aléatoire (uniforme ou de Mallows) des trames à chaque étape du processus de diffusion.

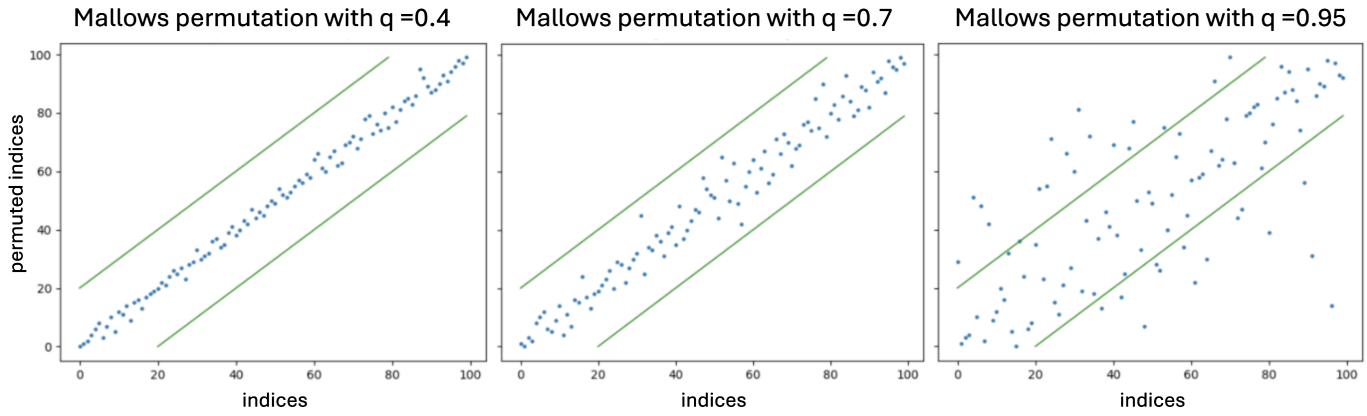

FIGURE 2 : Permutation de Mallows de paramètre q

2.2 RAVE

Une approche naïve d'édition de vidéos zero-shot consisterait à appliquer un modèle d'édition d'images à chaque trame, indépendamment des autres trames de la vidéo. L'expérience montre que cette solution est peu satisfaisante, car elle n'assure en rien la cohérence temporelle de la vidéo éditée.

RAVE [4] est un modèle d'édition zero-shot qui répond à ce problème grâce à une stratégie en grilles. Cette stratégie consiste à éditer un groupe d'images (typiquement une grille de 3×3) pour assurer la cohérence du style d'édition dans celles-ci. Dans le cas où les grilles regrouperaient des trames adjacentes de la vidéo, une cohérence temporelle serait simplement assurée par blocs de courte durée. En pratique, [4] induit une cohérence globale de la vidéo en reformant les grilles par tirage d'une permutation aléatoire uniforme au cours des itérations du processus de diffusion. Un schéma simplifié de l'architecture de [4] est présenté en Figure 1, où la permutation aléatoire appliquée est la permutation uniforme.

3 Méthode proposée

La permutation uniforme utilisée dans RAVE reforme les grilles en ordonnant les trames aléatoirement sans prendre en compte leur indice temporel. Ceci assure une bonne cohérence globale de la vidéo, mais généralement une mauvaise cohérence spatio-temporelle locale pour les vidéos longues ou présentant de forts mouvements de caméra. En effet, la concaténation de trames trop différentes biaise le processus de diffusion et provoque des phénomènes de *flickering* au niveau des détails de chaque trame. Pour pallier ce problème, nous proposons de tirer les permutations aléatoires de trames par une loi de Mallows [7] : une loi sur l'espace des matrices de permutation dont un paramètre (noté $q \in [0, 1]$) assure un certain contrôle en espérance sur la largeur de la bande des matrices générées.

3.1 Loi de Mallows

Mathématiquement, notons K le nombre d'images de la vidéo d'entrée et $\mathcal{S}_K$ l'ensemble des permutations sur $\mathcal{I} = \{1, \cdots, K\}$. Pour une permutation de Mallows $\pi \in \mathcal{S}_K$, on peut démontrer [7] qu'il existe une constante $c \in \mathbb{R}$ telle que :

$$c \cdot \min\{\lambda, K-1\} \leq \mathbb{E}|\pi(s) - s| \leq \min\{2\lambda, K-1\} \quad (1)$$

avec $\lambda = \frac{1}{1-q}$. Cette équation 1 implique que l'échantillonnage d'une loi de Mallows donne une permutation par bande où la largeur de la bande est controllée par le paramètre q. Lorsque $q \to 0$, les permutations générées se restreignent à des trames temporellement proches, et lorsque $q \to 1$, la bande des matrices de permutation générées s'élargit en moyenne. La figure 2 présente des permutations de Mallows pour différentes valeurs du paramètre q. Dans les cas limites, on retrouve respectivement la permutation identité ($q = 0$) et une permutation aléatoire uniforme ($q = 1$).

Concrètement, les permutations de Mallows sont tirées selon le processus décrit par l'algorithme 1. Les éléments de $\pi(\mathcal{I})$ sont ajoutés séquentiellement en suivant un processus d'insertion itératif. En partant du singleton contenant le premier élément de $\mathcal{I}$, le $k-$ème élément est ensuite inséré à

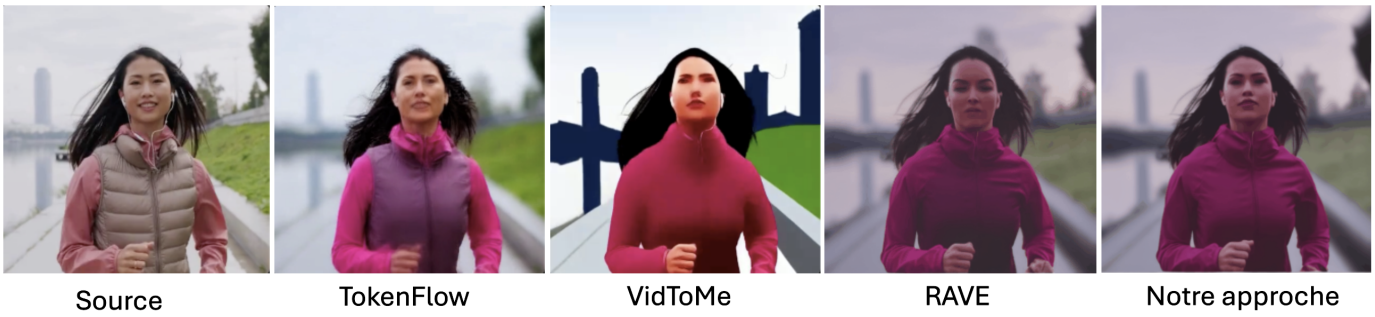

FIGURE 3 : Exemple qualitatif d'édition d'une vidéo DAVIS avec le prompt "*a woman is running at night*"

Algorithme 1 : Échantillonnage d'une permutation de Mallows

$\pi(\mathcal{I}) \leftarrow$ liste ordonnée $[0]$
pour $k = 1...K$ **faire**
  Tirer l'indice aléatoire i selon 2
  Insérer k à l'indice i dans $\pi(\mathcal{I})$
fin
return les indices permutés $\pi(\mathcal{I})$

l'indice $i \in \{0, 1, ..., k\}$ avec une probabilité :

$$p(k = i) = \frac{e^{-(1-q)(k-i)}}{\sum_{j=0}^{k} e^{-(1-q)(k-j)}}, \tag{2}$$

3.2 Adaptation de notre méthode

L'architecture de notre méthode est illustrée par la Figure 1, où le le bloc *Random shuffle* désigne l'utilisation de la loi de Mallows. Ainsi, à chaque étape du processus de débruitage, les trames de la vidéo sont permutées selon la loi de Mallows de paramètre q pour former les grilles de trames qui vont être passées dans le modèle d'édition d'images pré-entraîné.

L'hyperparamètre q de la permutation de Mallows permet de piloter le compromis souhaité entre la cohérence de l'édition et la préservation des détails lors du processus de diffusion. Pour une vidéo longue et mouvementée, une permutation locale (petit q) permettrait de bien conserver les détails de la vidéo d'entrée ; pour une vidéo courte et statique, une permutation globale (grand q) permettrait d'avoir un style général très uniforme. On retombe sur [4] pour $q = 1$.

4 Expériences

4.1 Protocole

L'évaluation de notre méthode suit le protocole proposés dans [4]. Les mêmes vidéos, extraites du jeu de données DAVIS [8], et les mêmes prompts ont été utilisés pour comparer quantitativement et qualitativement notre méthode avec celles de l'état de l'art.

Pour quantifier les résultats, trois métriques issues de [4, 12] sont évaluées : *Subject Consistency* mesure la persistence temporelle des objets entre trames, *CLIP-Frame* mesure la cohérence temporelle entre deux trames adjacentes, et *Warp-SSIM* mesure le respect du flow de la vidéo d'entrée.

TABLE 1 : Métriques d'évaluation sur le dataset de [4]

Method	**Subject consistency** ↑	**CLIP-F** ↑	**Warp-SSIM** ↑
TokenFlow [2]	0.9010	0.9635	0.4783
VidToMe [6]	0.8810	0.9693	**0.6494**
RAVE [4]	<u>0.9104</u>	<u>0.9709</u>	0.5012
Ours	**0.9125**	**0.9718**	<u>0.5251</u>

4.2 Méthodes comparées

La méthode proposée dans ce papier (avec $q = .9$) est comparée à trois modèles récents à la pointe en édition de vidéos zero-shot : TokenFlow [2], VidToMe [6] et RAVE [4].

Tous ces modèles utilisent le même *backbone* Stable Diffusion [10], ce qui permet bien de comparer les mérites de chaque architecture, indépendamment du modèle pré-entraîné de diffusion.

4.3 Résultats

L'exemple visuel de la Figure 3 permet une comparaison qualitative des différentes méthodes. On observe que TokenFlow et VidToMe ne respectent pas le prompt, contrairement à RAVE et notre méthode qui passent la scène sur des teintes sombres. VidToMe simplifie aussi fortement les textures, ce qui se traduit par une perte de réalisme. Notre méthode s'avère plus fidèle à la vidéo d'entrée que RAVE, qui change significativement les expressions du visage et ajoute un artefact derrière la tête du sujet.

En terme de métriques, la table 1 montre que notre méthode dépasse les *baselines* pour *Subject Consistency* et *CLIP-Frame*. Pour VidToMe, la simplification de textures assure un flow plus stable, donc plus proche de la vidéo source. Cependant, à niveau de réalisme équivalent, notre méthode a un score *Warp-SSIM* plus élevé que les *baselines*.

La Figure 4 illustre la stabilité temporelle des résultats obtenus par notre méthode appliquée à des vidéos de différentes longueurs avec une variété de prompts. On voit que notre méthode donne des résultats cohérents temporellement, avec une bonne gestion des occlusions et autres détails réalistes de la vidéo source. Ainsi, les expériences menées sur le jeu de données proposé par [4] tendent à montrer que notre méthode génère une qualité de résultats dépassant celle obtenue par d'autres méthodes de l'état de l'art pour l'édition de vidéos zero-shot.

FIGURE 4 : Exemples qualitatifs d’édition de style et de forme

5 Conclusion

Dans ce papier, nous avons proposé une méthode d’édition de vidéos zero-shot se basant sur un modèle de diffusion d’édition d’images à partir d’un texte d’entrée pré-entraîné. Les trames sont débruitées par grilles construites par une permutation de Mallows afin d’assurer une stabilité temporelle et préserver les détails de la vidéo source.

En choisissant la largeur de la permutation de Mallow, il est possible d’adapter notre méthode à la dynamique de la vidéo à traiter. Cependant, cet hyperparamètre est fixé pour l’ensemble de la vidéo. Une extension intéressante de notre approche consisterait à permettre une adaptation dynamique de ce paramètre en fonction de la quantité de mouvement observée localement à différents moments de la vidéo.

Références

[1] Duygu CEYLAN, Chun-Hao P HUANG et Niloy J MITRA : Pix2video : Video editing using image diffusion. *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23206–23217, 2023.

[2] Michal GEYER, Omer BAR-TAL, Shai BAGON et Tali DEKEL : Tokenflow : Consistent diffusion features for consistent video editing. *International Conference on Learning Representations*, 2024.

[3] Jonathan HO, Ajay JAIN et Pieter ABBEEL : Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[4] Ozgur KARA, Bariscan KURTKAYA, Hidir YESILTEPE, James M. REHG et Pinar YANARDAG : Rave : Randomized noise shuffling for fast and consistent video editing with diffusion models. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[5] Levon KHACHATRYAN, Andranik MOVSISYAN, Vahram TADEVOSYAN, Roberto HENSCHEL, Zhangyang WANG, Shant NAVASARDYAN et Humphrey SHI : Text2video-zero : Text-to-image diffusion models are zero-shot video generators. *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15954–15964, 2023.

[6] Xirui LI, Chao MA, Xiaokang YANG et Ming-Hsuan YANG : Vidtome : Video token merging for zero-shot video editing. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7486–7495, 2024.

[7] Colin L MALLOWS : Non-null ranking models. i. *Biometrika*, 44(1/2): 114–130, 1957.

[8] Federico PERAZZI, Jordi PONT-TUSET, Brian MCWILLIAMS, Luc VAN GOOL, Markus GROSS et Alexander SORKINE-HORNUNG : A benchmark dataset and evaluation methodology for video object segmentation. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 724–732, 2016.

[9] Chenyang QI, Xiaodong CUN, Yong ZHANG, Chenyang LEI, Xintao WANG, Ying SHAN et Qifeng CHEN : Fatezero : Fusing attentions for zero-shot text-based video editing. *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15932–15942, 2023.

[10] Robin ROMBACH, Andreas BLATTMANN, Dominik LORENZ, Patrick ESSER et Björn OMMER : High-resolution image synthesis with latent diffusion models. *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[11] Jiaming SONG, Chenlin MENG et Stefano ERMON : Denoising diffusion implicit models. *arXiv preprint arXiv :2010.02502*, 2020.

[12] Jiangshan WANG, Yue MA, Jiayi GUO, Yicheng XIAO, Gao HUANG et Xiu LI : Cove : Unleashing the diffusion feature correspondence for consistent video editing. *arXiv preprint arXiv :2406.08850*, 2024.

[13] Lvmin ZHANG, Anyi RAO et Maneesh AGRAWALA : Adding conditional control to text-to-image diffusion models, 2023.