Actes de CORIA-TALN 2026
Nantes, France, 29 juin au 3 juillet 2026

Revisite du légendage de différences d'images avec BLIP2IDC

Gautier Evennou[1,2], Antoine Chaffin[3], Vivien Chappelier[1,] Ewa Kijak[2]
(1) IMATAG, Rennes, France (2) IRISA, Univ. Rennes, INRIA, CNRS, Rennes, France
(3) LightOn, Paris, France * (4) Label4AI, Rennes, France*
{prenom.nom}@irisa.fr

RÉSUMÉ

L'amélioration des modèles génératifs permet la génération d'images altérées à grande échelle. Pour contrer les mésusages de cette technologie, la tâche d'*Image Difference Captioning* (IDC) vise à décrire les différences entre deux images. Bien que cette tâche *image2text* soit traitée avec succès pour des images 3D simples, elle rencontre des difficultés avec des images du monde réel. Cela est dû à la rareté des données d'entraînement et à la difficulté de capturer des différences fines entre des images complexes. Pour répondre à ces défis, nous proposons dans cet article un cadre simple mais efficace pour adapter les modèles existants de légendage d'images à la tâche IDC et pour enrichir les jeux de données IDC. Nous introduisons *BLIP2IDC*, une adaptation de BLIP2 à la tâche IDC à faible coût computationnel, et montrons qu'il surpasse les approches existantes de manière significative sur les jeux de données IDC d'images réelles. Nous proposons également d'utiliser l'augmentation synthétique pour améliorer les performances des modèles IDC de manière agnostique. Nous démontrons que cette stratégie fournit des données de haute qualité, conduisant à un nouveau jeu de données exigeant et bien adapté à l'IDC, nommé Syned.

ABSTRACT

Reframing Image Difference Captioning with BLIP2IDC and Synthetic Augmentation

The rise of the generative models quality during the past years enabled the generation of edited variations of images at an important scale. To counter the harmful effects of such technology, the Image Difference Captioning (IDC) task aims to describe the differences between two images. While this *image2text* task is successfully handled for simple 3D rendered images, it struggles on real-world images. The reason is twofold : the training data-scarcity, and the difficulty to capture fine-grained differences between complex images. To address those issues, we propose a simple yet effective framework to both adapt existing image captioning models to the IDC task and augment IDC datasets. We introduce BLIP2IDC, an adaptation of BLIP2 to the IDC task at low computational cost, and show it outperforms two-streams approaches by a significant margin on real-world IDC datasets. We also propose to use synthetic augmentation to improve the performance of IDC models in an agnostic fashion. We show that our strategy provides high quality data, leading to a challenging new dataset well-suited for IDC named Syned.

MOTS-CLÉS : Légendage de différences d'image, Perturbations sémantiques, Augmentation de données, jeu de données IDC.

KEYWORDS: Image difference captioning, Semantic Perturbations, Synthetic augmentation, IDC dataset.

∗. Travaux faits à IMATAG/IRISA.

Actes de l'atelier sur l'évaluation des modèles génératifs (LLM) et challenges (EvalLLM 2026), pages 160–178.

1 Introduction

La désinformation, souvent propagée par des images manipulées ou sorties de leur contexte, représente un défi majeur. L'*Image Difference Captioning* (IDC) offre une solution en générant des descriptions textuelles permettant aux humains de déterminer facilement si une image a subi des altérations sémantiques. Cet article explore l'IDC (Park *et al.*, 2019), une approche récente qui va au-delà de l'analyse traditionnelle d'images en produisant des descriptions textuelles détaillées des différences entre deux images. L'IDC trouve des applications dans divers domaines, allant de la détection de changements subtils dans les images satellites (Jhamtani & Berg-Kirkpatrick, 2018; Chouaf *et al.*, 2021; Qiu *et al.*, 2021; Liu *et al.*, 2023a) à l'identification d'anomalies dans l'imagerie médicale (Liu *et al.*, 2021) ou à l'explication de la manipulation des images.

Bien que légender une seule image présente déjà des défis considérables en soi, l'IDC introduit des complexités supplémentaires, car elle implique de décrire les différences subtiles présentes dans une paire d'images similaires. Idéalement, les légendes devraient ignorer les objets communs entre les images et se concentrer sur les changements nuancés entre elles. Les progrès de l'IDC ces dernières années reposent en grande partie sur l'émergence des modèles vision-langage et des techniques d'apprentissage inter-domaines.

Le deuxième défi de l'*IDC* est la disponibilité d'un jeu de données suffisamment grand et diversifié pour cette tâche. La création d'un jeu de données *IDC* de haute qualité est particulièrement difficile et coûteuse en ressources, car elle nécessite des paires d'images accompagnées de descriptions détaillées de leurs différences, couvrant divers types de modifications. Ce processus peut reposer sur un travail externalisé coûteux et chronophage (Zhang *et al.*, 2023), ou alternativement, sur l'utilisation de scènes rendues en 3D (Park *et al.*, 2019), ou sur la capture de variations temporelles (Tan *et al.*, 2019). Les pipelines existants d'entraînement et d'évaluation pour l'*IDC* présentent des lacunes telles que des métriques sous-optimales, des incohérences dans les vérités terrain, une taille réduite des jeux de données réels, ou l'absence de catégorisation des modifications.

Dans cet article, nous proposons une utilisation innovante des données synthétiques et des architectures multimodales avancées pour répondre aux limitations de l'*IDC* en termes de données et de modèles. Nous proposons l'application de modèles génératifs (Brooks *et al.*, 2023; Rombach *et al.*, 2021) pour

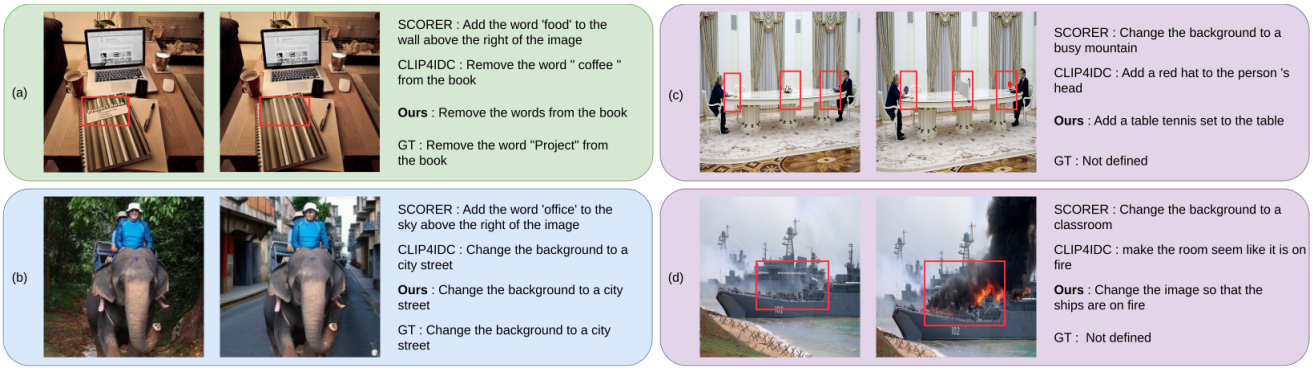


FIGURE 1 – Résultats des modèles IDC sur : (a) un échantillon d'entraînement du jeu de données Emu Edit, (b) un échantillon de test et (c-d) des échantillons en *zero-shot* issus du web. **BLIP2IDC** (notre approche) est capable de capturer des différences fines, de décrire des scènes complexes et de bien généraliser à des données inédites.

produire des données synthétiques, fournissant aux modèles *IDC* des paires d'images diverses et complexes. Concernant les avancées dans les modèles multimodaux, nous discutons de la manière dont le pré-entraînement de modèles multimodaux comme CLIP (Guo *et al.*, 2022) et BLIP2 (Li *et al.*, 2023) est déterminant pour les tâches d'*IDC*. Nous explorons le potentiel de BLIP2 dans l'*IDC*, offrant une solution plus « prête à l'emploi » qui contraste fortement avec la dépendance des modèles précédents à des entraînements complexes, multi-étapes et à des processus séparés pour le codage des images. Malgré sa taille plus importante et les coûts associés au fine-tuning, nous démontrons la faisabilité et l'efficacité de LoRa (Hu *et al.*, 2022) pour adapter BLIP2 à l'*IDC*, en capitalisant sur les vastes capacités de ce modèle tout en maintenant un coût d'adaptation faible. Nos principales contributions [1] sont les suivantes :

1. Nous fournissons un cadre pour l'augmentation synthétique basés sur des modèles de diffusion afin de concevoir des jeux de données *IDC* adaptés. Nous publions **Syned**, notre version augmentée synthétiquement de Emu Edit (Sheynin *et al.*, 2023), un jeu de données d'édition d'images, pour fournir un benchmark exigeant pour la tâche *IDC*. Nous démontrons l'amélioration significative apportée par cette augmentation synthétique ciblée.
2. Nous proposons **BLIP2IDC**, un nouveau modèle *IDC* basé sur BLIP2. Des expériences approfondies menées sur des jeux de données synthétiques et réels démontrent les fortes performances de BLIP2IDC dans des scénarios réels, ainsi qu'une bonne généralisation à de nouvelles données (Fig. 1).
3. Nous présentons une évaluation complète de plusieurs modèles état de l'art sur un nouveau jeu de données *IDC* réel basé sur les sorties de Emu Edit, contribuant à une meilleure compréhension des capacités actuelles des modèles en *IDC*.

2 Travaux connexes

Légendage d'images et modèles multimodaux. Le légendage des différences entre images (*Image Difference Captioning*) est étroitement lié au légendage d'images (*Image Captioning*, IC) et à la réponse aux questions visuelles (*Visual Question Answering*, VQA). Le légendage d'images (Li *et al.*, 2022; Wang *et al.*, 2022; Li *et al.*, 2023) vise à décrire le contenu d'une image avec des légendes fines. Les modèles IC sont entraînés sur des jeux de données à l'échelle du web (Lin *et al.*, 2014; Hodosh *et al.*, 2013; Young *et al.*, 2014) pour exploiter autant de connaissances visuelles que possible. Les modèles IC tentent de connecter le texte et les images avec diverses solutions : arbres syntaxiques à partir de caractéristiques d'images (Mitchell *et al.*, 2012), décodage RNN (Hochreiter & Schmidhuber, 1997) des caractéristiques CNN avec (Xu *et al.*, 2015) ou sans (Vinyals *et al.*, 2015; Bahdanau *et al.*, 2015) mécanisme d'attention, tâches de pré-entraînement utilisant des ancres basées sur des objets (Li *et al.*, 2020), apprentissage Seq2Seq avec un vocabulaire unifié pour tous les jetons linguistiques et visuels (Wang *et al.*, 2022), ou encore l'utilisation d'intégrations supplémentaires pour établir un pont entre les jetons visuels et textuels (Li *et al.*, 2023). Les modèles multimodaux spécialisés dans le VQA utilisent l'attention inter-modale pour ajuster l'extraction des caractéristiques visuelles en fonction de l'invite textuelle, permettant ainsi une interaction et focalisant le modèle sur les informations pertinentes pour la question. Bien que très efficaces (Wang *et al.*, 2024; Liu *et al.*, 2023b), ces modèles sont sujets à des incohérences en raison de leur dépendance à l'invite (Zhou *et al.*, 2022).

1. Le code et les données sont disponibles sur https://github.com/gautierevn/BLIP2IDC

Légendage des différences entre images (*Image Difference Captioning*, IDC). L'*IDC* se concentre sur la distinction des écarts entre deux images quasi identiques. Les différences modifiant le sens de l'image (changements sémantiques) sont celles sur lesquelles le modèle doit se concentrer. À l'inverse, les changements éditoriaux (non sémantiques) tels que la compression ou le redimensionnement doivent être ignorés. Les travaux antérieurs rencontrent deux difficultés majeures : comment représenter les caractéristiques des différences et comment collecter ce type spécifique de données. Les architectures basées sur l'apprentissage profond, telles que les CNN, les encodages basés sur CLIP et les RNN, sont largement utilisées dans ce domaine pour apprendre les caractéristiques (Jhamtani & Berg-Kirkpatrick, 2018; Shi *et al.*, 2020a; Hosseinzadeh & Wang, 2021; Yao *et al.*, 2022; Qiu *et al.*, 2021; Chouaf *et al.*, 2021; Liu *et al.*, 2023a). Les travaux précédents extraient les caractéristiques de chaque image indépendamment, négligeant ainsi la corrélation entre les images dans l'espace des pixels. Après cette étape d'extraction, les caractéristiques sont fusionnées par la concaténation (Tu *et al.*, 2023c; Yao *et al.*, 2022; Qiu *et al.*, 2021; Park *et al.*, 2019) avant d'être fournies soit à un encodeur de différences pour la représentation des changements, soit directement à un décodeur pour générer les descriptions textuelles. L'encodeur et le décodeur de différences sont basés sur des transformers (Vaswani *et al.*, 2017) ou des RNN (Qiu *et al.*, 2021). Un autre approche consiste à définir des tâches de pré-entraînement pour trouver un espace de représentation adapté aux différences (Tu *et al.*, 2023c). Récemment, l'utilisation d'un modèle multimodal tel que CLIP (Radford *et al.*, 2021) permet une meilleure représentation, atteignant ainsi de meilleures performances. Cependant, l'adaptation actuelle à l'*IDC* (Guo *et al.*, 2022) peine à exploiter pleinement son pré-entraînement. De plus, comme les données de haute qualité sont difficiles à trouver, la plupart des méthodes sont entraînées et évaluées sur des jeux de données rendus en 3D, qui ne représentent pas avec précision les performances en conditions réelles.

Édition d'images. Les progrès récents dans les modèles génératifs (Goodfellow *et al.*, 2014; Rombach *et al.*, 2021) permettent une génération réaliste d'images. Cela stimule davantage les travaux (Meng *et al.*, 2021; Brooks *et al.*, 2023; Parmar *et al.*, 2023) sur l'édition d'images, où des modifications sémantiques sont apportées à une image à l'aide d'une invite textuelle, tout en conservant le sujet principal de l'image. Les travaux récents (Sheynin *et al.*, 2023; Zhang *et al.*, 2023) se sont concentrés sur l'utilisation de jeux de données texte-image existants du monde réel, tels que MSCOCO (Lin *et al.*, 2014), pour effectuer des éditions sur ces images en s'appuyant sur des modifications de la légende associée. Le principal obstacle de ces méthodes réside dans le processus de génération et de vérification entièrement supervisé, où chaque génération est examinée par des opérateurs et plusieurs niveaux de filtres sont utilisés pour conserver les meilleures générations. Bien que ce processus produise un jeu de données de haute qualité, il est chronophage et coûteux. Nous montrons dans cet article que cette configuration permet néanmoins la création automatique de jeux de données pour l'*IDC* : un couple est créé en utilisant l'image originale et sa version modifiée, et l'invite *(prompt)* utilisée pour effectuer l'édition sert de description de la modification cible.

3 BLIP2IDC

BLIP2 (Li *et al.*, 2023) est un modèle multimodal qui introduit le *Querying Transformer* (QFormer) comme bloc principal pour connecter les représentations texte et image. Il repose sur un pré-entraînement en deux étapes : une étape d'apprentissage de la représentation vision-langage avec un encodeur d'images figé, et une étape d'apprentissage génératif vision-vers-langage avec un modèle de langage (LLM) figé. Le QFormer est composé de deux modules transformers. Le premier

s'intéresse à l'image via une attention croisée entre les sorties de l'encodeur visuel et des plongements appris, appelées requêtes (*queries*). Le second module interagit avec le texte de référence et avec les requêtes pour garantir l'alignement vision-langage. BLIP2 est pré-entraîné pour optimiser conjointement trois objectifs de pré-entraînement : l'*apprentissage contrastif image-texte*, la *correspondance image-texte*, et la *génération de texte ancrée sur l'image*. Dans cet article, nous montrons que, en tirant parti de son pré-entraînement pour le légendage d'images, BLIP2 peut être adapté à l'*IDC* à un coût computationnel faible et sans modifier son architecture.

3.1 Adaptation

Nous soutenons que le schéma classique d'encodage à deux flux pour l'*IDC* (Fig. 2, haut) conduit à une comparaison sous-optimale des images. Pour le légendage d'images standard, BLIP2 prend une seule image en entrée et utilise un modèle ViT (Dosovitskiy *et al.*, 2021; Vaswani *et al.*, 2017) comme encodeur d'images figé. Dans le contexte de l'*IDC*, alimenter BLIP2 simultanément avec les deux images à comparer permet aux couches d'attention de l'encodeur visuel et du QFormer de se concentrer précocement sur les différences entre les deux images, et d'encoder les deux images en relation l'une avec l'autre plutôt que séparément.

Dans notre pipeline d'adaptation de BLIP2 (Fig. 2, bas), nous fournissons une seule image en entrée, résultant de la concaténation verticale des deux images, permettant au modèle de prêter attention aux différences dès le début, tout en évitant toute modification de l'architecture du modèle. Contrairement à BLIP2, qui n'entraîne que le QFormer, le ViT et le LLM doivent être ajustés (*fine-tuned*) pour compenser cette modification du domaine d'entrée. Bien que les images soient étirées du fait de la concaténation, cela ne nuit pas aux performances. Nous émettons l'hypothèse que ce comportement est dû à la manière dont BLIP2 a été pré-entraîné, dans laquelle les images sont recadrées aléatoirement puis étirées à une taille carrée fixe, sans remplissage. Ainsi, le modèle BLIP2 a appris à être robuste à diverses opérations d'étirement.

3.2 Fine-tuning efficace

Fine-tuning des modules. Bien que BLIP2IDC ne nécessite pas de modifier l'architecture du modèle, celui-ci doit être ajusté pour la tâche *IDC* afin de s'adapter à la nouvelle tâche et au type de données. Contrairement à l'entraînement original de BLIP2, qui n'entraine que le QFormer, tous les composants, y compris le ViT, le QFormer et le décodeur de texte ancré sur l'image, sont ajustés pour obtenir les meilleures performances en *IDC* (voir la Fig. 5).

Low Rank Adaptation (LoRA). Nous ajustons par LoRA pour réduire les ressources d'entraînement tout en maintenant les performances. En ajustant seulement 0,1% de tous les paramètres – spécifiquement les couches Q, K, V des modules d'attention – nous atteignons des performances de pointe avec une utilisation minimale de ressources. Cette approche permet à BLIP2IDC de s'adapter aux tâches *IDC*, en équilibrant bonne performance et efficacité des ressources.

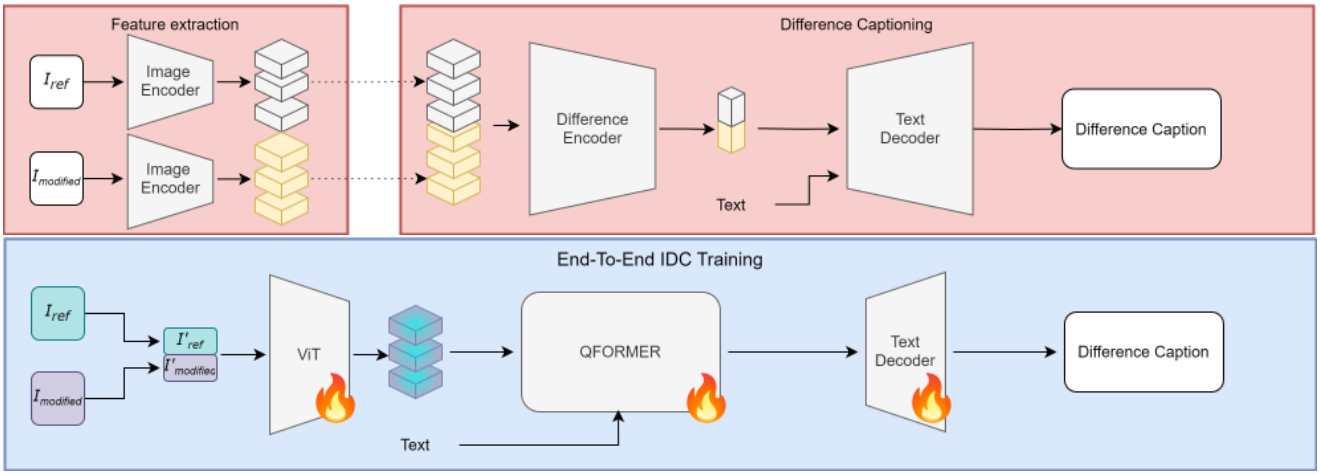


FIGURE 2 – En haut : le pipeline classique de l'*IDC*. L'étape d'extraction des caractéristiques est réalisée avant l'entraînement avec un encodeur d'images figé. Lors de l'apprentissage, la représentation des différences est apprise à partir de la concaténation des caractéristiques extraites des deux images. La sortie est ensuite fournie au décodeur de texte pour effectuer la génération de la légende des différences de manière autorégressive. En bas : le pipeline *end-to-end* de BLIP2IDC. Les images sont redimensionnées et concaténées avant d'être fournies à l'architecture BLIP2, ajustée par LoRA.

3.3 Avantages par rapport aux autres modèles IDC

Les modèles les plus récents pour l'*IDC* sont CLIP4IDC (Guo *et al.*, 2022) et SCORER (Tu *et al.*, 2023c). CLIP4IDC utilise une approche d'entraînement en deux étapes : d'abord pour adapter les représentations visuelles des paires d'images pour le légendage, puis pour générer des légendes basées sur l'encodage des différences visuelles, respectivement à travers des pertes contrastives et d'entropie croisée. SCORER suit le cadre traditionnel de l'*IDC* tel qu'illustré dans la Figure 2. Il améliore les performances avec des modules innovants pour l'invariance aux changements de vue et l'informativité des légendes, et est présenté comme montrant une force particulière sur le jeu de données CLEVR-DC.

Ces modèles se caractérisent par des procédures d'entraînement complexes avec plusieurs étapes et des processus d'encodage d'images séparés. En revanche, notre modèle BLIP2IDC concatène les images pour un encodage conjoint, en utilisant des mécanismes d'attention précoces tout en apprenant en une seule étape. Cette approche s'avère cruciale pour l'*IDC*, comme en témoignent nos résultats expérimentaux dans la Section 5. BLIP2IDC bénéficie également des connaissances acquises grâce au pré-entraînement non supervisé à grande échelle de BLIP2 pour le légendage d'images, ce qui le rend attrayant compte tenu des défis associés à l'entraînement complet (*from scratch*) d'un modèle *IDC* avec peu de données.

4 Jeux de données

La structure des jeux de données *IDC* est un triplet $(I_{ref}, I_{modified}, GT)$ qui représente respectivement l'image originale I_{ref}, l'image modifiée $I_{modified}$ et l'ensemble des descriptions de référence des différences GT.

4.1 Jeux existants

Nous comparons notre approche aux modèles *IDC* récents sur des jeux de données standards : CLEVR-Change (Park *et al.*, 2019), CLEVR-DC (Kim *et al.*, 2021), Spot-The-Diff (Qiu *et al.*, 2021) et Image Editing Request (Tan *et al.*, 2019). Ces jeux de données se distinguent par diverses propriétés, résumées dans le Tableau 1 : le nombre de paires d'images $(I_{ref}, I_{modified})$, l'origine des images I_{ref}, qui peuvent être des images réelles ou générées par des modèles 3D, le nombre de légendes de référence (vérités terrain GT) par triplet, la manière dont $I_{modified}$ et GT sont obtenus, le nombre et le type de transformations effectuées, et si une intervention manuelle a été nécessaire pour créer le jeu de données. Cela peut se produire au niveau du filtrage des triplets ou de la création des légendes de référence.

Ces propriétés peuvent limiter la création de ces jeux de données, qui peut être coûteuse, ainsi que leur utilité, par exemple lorsque la diversité des types de transformations est réduite. Nous discutons des limitations de ces jeux de données en annexe B, ainsi que des problèmes posés par l'évaluation de la tâche *IDC*. Pour remédier à ces limitations, nous proposons une méthode basée sur des méthodes d'édition d'images guidées par du texte pour créer des jeux de données *IDC* sans intervention manuelle, qui peuvent ensuite être adaptés à différents cas d'usage.

Par ailleurs, les capacités d'édition d'images basées sur des instructions sont évaluées sur certains benchmarks (jeux de données IE), comme MagicBrush (Zhang *et al.*, 2023) ou les générations de l'ensemble de test Emu Edit (EE) (Sheynin *et al.*, 2023). Ces jeux de données sont également composés de triplets $(I_{ref}, T_{instruction}, I_{edited})$, où $T_{instruction}$ est la modification qui transforme I_{ref} en I_{edited} et I_{edited} est l'image modifiée. Nous proposons donc d'adapter ces jeux de données pour la tâche *IDC*. Tous ces jeux de données sont décrits en annexe A.

TABLE 1 – Jeux de données *IDC* et *IE*. *Setting* indique l'origine des images de référence, qu'elles proviennent de scènes 3D ou du monde réel. La *méthode d'édition* fait référence à la manière dont les images modifiées sont créées ou collectées. *GT* correspond au nombre de vérités terrain dans les ensembles de validation et de test.

Jeu de données	# paires d'image	setting	méthode d'édition	curation	# GT
CLEVR-Change (Park *et al.*, 2019)	79,606	Scènes 3D Scenes	Moteur 3D	✓	5.0
CLEVR-DC (Kim *et al.*, 2021)	48,000	Scènes 3D	Moteur 3D	✓	5.0
STD (Qiu *et al.*, 2021)	13,192	Real-world	Changement temporel		1.86
IER (Tan *et al.*, 2019)	3,939	Real-world	Edition humaine	✓	3.0
MagicBrush (Zhang *et al.*, 2023)	10,308	Real-world	DallE-2 (Ramesh *et al.*, 2022) platform	✓	1.0
EE (Sheynin *et al.*, 2023)	5,612	Real-world	Génération automatique	✓	1.0
Syned(ours)	28,720	Real-world	Génération automatique		5.0

4.2 Augmentation artificielle

Pour pallier les lacunes actuelles des ensembles de données IDC mentionnés ci-dessus, nous proposons un pipeline (Fig. 3) permettant de générer des échantillons d'apprentissage synthétiques à partir d'images originales réelles. L'idée est d'exploiter les nouveaux modèles d'édition d'images promptables afin de produire des images modifiées à partir d'un ensemble d'images originales et d'instructions d'édition relevant d'une gamme définie de modifications. Des légendes de référence supplémentaires sont générées par un LLM comme variations des instructions d'édition, garantissant ainsi des vérités terrain cohérentes. Un tel pipeline peut être utilisé soit pour enrichir un jeu de

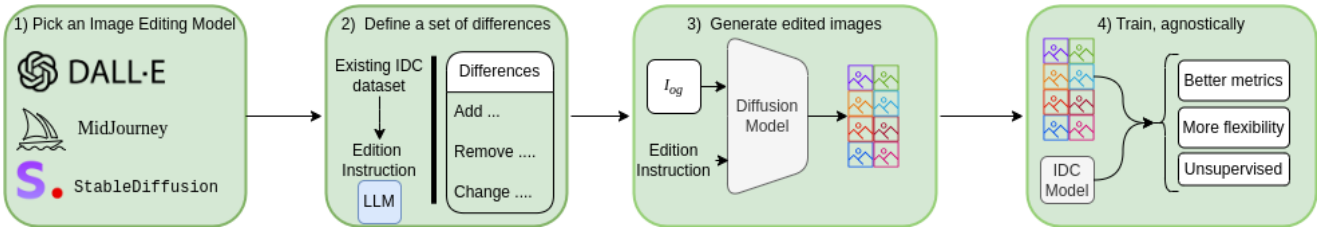


FIGURE 3 – Pipeline de création du jeu de données synthétique exploitant un LLM et un modèle d'édition promptable.

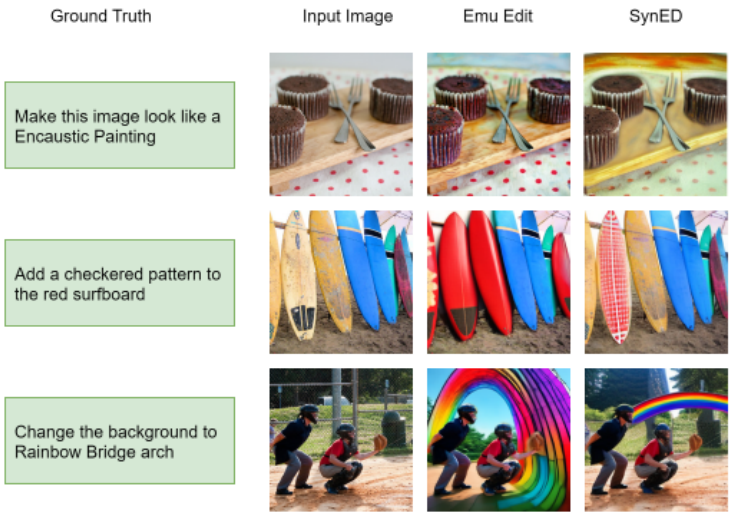


FIGURE 4 – Comparaison d'échantillons du jeu d'entrainement d'EE, générés par Emu Edit, et de Syned, générés par un modèle InstructPix2Pix ajusté.

données *IDC* existant, soit pour créer un nouveau jeu de données dont les images et les modifications sont adaptées à une tâche spécifique en aval avec de nouveaux types de données ou de modifications.

Pour évaluer ce pipeline, nous l'avons appliqué pour enrichir le jeu de données EE (Sheynin *et al.*, 2023), car ce jeu de données fournit des images éditées du monde réel accompagnées d'instructions d'édition bien adaptées à un modèle d'édition d'images, avec une distinction claire entre chaque classe de modifications. Comme modèle d'édition d'images, nous avons choisi le modèle InstructPix2Pix (Brooks *et al.*, 2023) ajusté sur le jeu de données MagicBrush (Zhang *et al.*, 2023). L'ensemble d'entraînement est enrichi en générant 8 nouvelles images modifiées pour chaque image originale du jeu de données EE, ainsi que leurs instructions d'édition associées, sans nettoyage manuel.

Pour garantir une évaluation significative, nous utilisons le modèle Llama-2-7b-chat-hf pour générer 4 variations supplémentaires de chaque instruction d'édition de l'ensemble de test EE. Chacun des 2 022 échantillons de l'ensemble de test est donc décrit par 5 légendes de référence : l'instruction originale et les 4 légendes de référence fournies par le LLM. Notons que les images de l'ensemble d'entraînement (voir Fig. 4) n'ont pas été modifiées par le même modèle d'édition que celui utilisé pour l'ensemble de test.

Le jeu de données Syned résultant comprend :

— **un ensemble d'entraînement de 28 720 paires** $(I_{ref}, I_{modified})$: 8 variations de chacune

des 3 590 images originales de MSCOCO, générées par InstructPix2Pix. Les différentes modifications sont catégorisées en 8 classes.
— **un ensemble de test de 2 022 échantillons avec 5 références chacun**, où les images modifiées ont été générées par le modèle Emu Edit.

Bien que les modifications puissent être définies selon les besoins spécifiques, elles doivent être alignées avec les capacités du modèle d'édition d'images. Par exemple, les modèles d'édition actuels ne parviennent pas à ajouter du texte aux images. De telles instructions introduisent de mauvais échantillons dans l'ensemble d'entraînement. À mesure que les modèles d'édition d'images s'améliorent, le pipeline proposé permettra d'inclure des éditions plus fiables dans l'ensemble d'entraînement des modèles *IDC*. Il permet néanmoins déjà une génération personnalisée de jeux de données *IDC* plus exigeants.

5 Expériences

5.1 Protocole d'évaluation

Modèles et jeux de données. Nous comparons BLIP2IDC à d'autres modèles *IDC* de pointe avec les résultats rapportés dans leurs articles respectifs sur les quatre jeux de données *IDC* standards : CLEVR-Change, CLEVR-DC, STD et IER. Nous introduisons également le jeu de données EE dans la tâche *IDC*, sur lequel nous entraînons et évaluons également CLIP4IDC (Guo *et al.*, 2022) et SCORER (Tu *et al.*, 2023c) à des fins de comparaison.

Fine-tuning. BLIP2IDC est ajusté comme décrit dans la Section 3.2, avec un LoRA de rang 8. Pour notre propre adaptation de BLIP2, nous utilisons la version LLM de BLIP2 avec décodeur uniquement, avec opt2.7B et vit-base-patch16-224 comme modèle ViT. Nous utilisons l'implémentation de LoRA de la bibliothèque Python peft (Mangrulkar *et al.*, 2022) pour entraîner BLIP2 et stocker l'ajustement des poids dans un fichier léger de 20 Mo. Pour CLIP4IDC et SCORER, nous avons utilisé soit les hyperparamètres de leurs dépôts en ligne respectifs, soit les hyperparamètres communiqués par les auteurs.

Augmentation de données standard. Pour augmenter la robustesse du modèle, nous utilisons *RandomGaussianBlur* et la *compression JPEG* comme schéma d'augmentation de données non perturbateur. Les augmentations doivent en effet laisser le contenu sémantique de la vérité terrain intact. Les transformations basées sur le recadrage, le jitter de couleur, le retournement horizontal/vertical, la luminosité ou le contraste peuvent en effet ajouter des différences significatives et ainsi modifier la légende de différence attendue.

Métriques. Sur la base des travaux précédents (Jhamtani & Berg-Kirkpatrick, 2018; Shi *et al.*, 2020a; Hosseinzadeh & Wang, 2021; Yao *et al.*, 2022; Qiu *et al.*, 2021; Chouaf *et al.*, 2021; Liu *et al.*, 2023a), le CIDEr (C) est la métrique principale utilisée pour évaluer les légendes de différences générées. Cette approche basée sur les n-grammes garantit que le texte évalué capture les aspects les plus pertinents de l'image, tels que convenus par plusieurs annotateurs humains. Dans un contexte dépendant de la sémantique, la métrique CIDEr récompense les descriptions qui reflètent avec précision la compréhension consensuelle du contenu de l'image. BLEU-4 (B), ROUGE-L (R) et METEOR (M) sont également utilisés comme métriques secondaires pour évaluer la qualité des phrases.

Méthode	Couleur	Texture	Déplacement	Ajout	Suppression
DUDA (Park *et al.*, 2019) (ICCV 2019)	120.4	86.7	56.4	108.2	103.4
VAM+ (Shi *et al.*, 2020b) (ECCV 2020)	122.1	98.7	82.0	126.3	115.8
IFDC (Huang *et al.*, 2022)(TMM 2022)	133.2	99.1	82.1	128.2	118.5
DUDA+ (Hosseinzadeh & Wang, 2021) (CVPR 2021)	120.8	89.9	62.1	119.8	123.4
BiDiff (Sun *et al.*, 2022) (IJIS 2022)	115.9	106.8	71.8	121.3	124.9
IDC-PCL (Yao *et al.*, 2022) (AAAI 2022)	131.2	101.1	81.7	133.3	116.5
CLIP4IDC (Guo *et al.*, 2022) (ACL-IJCNLP 2022)	149.1	135.3	91.0	132.4	**135.5**
SCORER (Tu *et al.*, 2023c) (ICCV 2023)	143.2	135.2	91.6	129.4	132.6
SCORER+CBR (Tu *et al.*, 2023c)(ICCV 2023)	146.2	133.7	92.2	131.1	133.9
BLIP2IDC (le nôtre)	**152.31**	**137.0**	**99.8**	**135.6**	133.1

TABLE 2 – Scores CIDEr sur CLEVR-Change pour les 5 catégories de changement.

5.2 Évaluation de BLIP2IDC

Nous évaluons les performances de BLIP2IDC par rapport à d'autres méthodes sur les quatre jeux de données *IDC* existants. Tous les résultats, à l'exception de CLIP4IDC sur CLEVR-DC et BLIP2IDC, sont rapportés à partir de travaux précédents. Les résultats selon différentes configurations de BLIP2IDC, en fonction des modules ajustés, sont donnés en annexe C.

Résultats sur les scènes 3D. Les résultats par type de changement sémantique sur CLEVR-Change sont donnés dans la Table 2. BLIP2IDC se classe en première position pour presque tous les types de changements sémantiques, avec une augmentation de 10% des performances de légendage sur le changement le plus difficile « Déplacement », et une amélioration significative par rapport à la méthode de pointe précédente sur la métrique CIDEr. Sur CLEVR-DC, qui introduit des changements extrêmes de points de vue, nous observons que CLIP4IDC obtient les meilleurs résultats, suivi de BLIP2IDC (Table 3). Ces deux modèles surpassent SCORER, bien que ce dernier ait été spécialement conçu pour garantir l'invariance aux changements extrêmes de point de vue.

Résultats sur les images du monde réel.

Selon la Table 3, la méthode BLIP2IDC se classe en première position sur la métrique CIDEr avec une marge significative de 9,2% par rapport à l'état de l'art sur STD. Le jeu de données IER est composé d'images retouchées avec Photoshop dans un cadre réel. Ainsi, l'amélioration impressionnante par rapport à l'état de l'art précédent est attendue, car BLIP2 a été pré-entraîné sur 129 millions d'images rélles (Lin *et al.*, 2014; Krishna *et al.*, 2017; Sharma *et al.*, 2018). Nos résultats montrent que notre adaptation surpasse les modèles existants, même ceux basés sur CLIP, qui a également été entraîné à grande échelle. Dans l'ensemble, ces résultats montrent qu'une adaptation efficace d'un modèle pré-entraîné puissant permet d'atteindre des performances de pointe, alors que la plupart des modèles existants sont moins performants en raison du manque de données pendant l'entraînement.

5.3 Jeu de données généré automatiquement

Nous avons introduit dans la Section 4.2 Syned, un nouveau jeu de données conçu pour la tâche *IDC*, basé sur le jeu de données EE. Les nouvelles données sont générées par un modèle différent d'édition d'images basé sur des instructions, élargissant ainsi la portée et la diversité des nouvelles images modifiées par rapport au jeu de données EE original (voir Fig. 4). Nous évaluons l'utilité de Syned comme augmentation de données pour différents modèles *IDC* dans la Table 4 qui montre une

Méthode	CLEVR-DC B	CLEVR-DC M	CLEVR-DC C	STD B	STD M	STD C	IER B	IER M	IER R	IER C
Dyn rel-att (Tan *et al.*, 2019)	-	-	-	-	-	-	6.7	12.8	37.5	26.4
M-VAM (Shi *et al.*, 2020b)	40.9	27.1	60.1	10.1	12.4	38.1	-	-	-	-
M-VAM+RAF (Shi *et al.*, 2020b)	-	-	-	11.1	12.9	43.5	-	-	-	-
VA (Kim *et al.*, 2021)	44.5	29.2	70.0	-	-	-	-	-	-	-
VACC (Kim *et al.*, 2021)	45.0	29.3	71.7	9.7	12.6	41.5	-	-	-	-
DUDA (Park *et al.*, 2019)	40.3	27.1	56.7	8.1	12.5	34.5	6.5	12.4	37.3	22.8
NCT (Tu *et al.*, 2023b)	47.5	32.5	76.9	-	-	-	8.1	15.0	38.8	34.2
SRDRL+AVS (Tu *et al.*, 2021)	-	-	-	-	13.0	35.3	-	-	-	-
IFDC (Huang *et al.*, 2022)	-	-	-	8.7	11.7	37.0	-	-	-	-
BDLSCR (Sun *et al.*, 2022)	-	-	-	6.6	10.6	42.2	6.9	14.6	38.5	27.7
VARD-Trans (Tu *et al.*, 2023a)	48.3	32.4	77.6	-	12.5	30.3	10.0	14.8	39.0	35.7
MCCFormers-D (Qiu *et al.*, 2021)	46.9	31.7	71.6	10.0	12.4	43.1	8.3	14.3	39.2	30.2
SCORER (Tu *et al.*, 2023c)	49.5	**33.4**	82.4	9.4	13.8	38.5	9.6	14.6	39.5	31.0
SCORER+CBR (Tu *et al.*, 2023c)	49.4	**33.4**	83.7	10.2	12.2	38.9	10.0	15.0	39.6	33.4
CLIP4IDC (Guo *et al.*, 2022)	**54.7**†	33.0†	**89.9**†	**11.6**	**14.2**	47.4	8.2	14.6	40.4	32.2
BLIP2IDC (la nôtre)†	49.3	33.0	88.5	11.4	13.5	**51.4**	**17.4**	**20.1**	**48.5**	**74.1**

TABLE 3 – Résultats sur CLEVR-DC, STD, et IER (rapportés de leurs articles correspondants), sauf † rapportés de notre propres expériences.

Modèles	**EE**	**Syned+ EE**	Amélioration (%)
SCORER	21.1	23.2	-
CLIP4IDC	32.4	35.9	10.8
BLIP2IDC	100.0	106.8	6.8

TABLE 4 – Scores CIDEr sur le jeu de test Emu Edit montrant des améliorations constantes avec l'augmentation synthétique (EE + Syned) sur tous les modèles.

amélioration avec l'augmentation synthétique proposée (EE + Syned) pour CLIP4IDC et BLIP2IDC. Notons que l'implémentation disponible de SCORER permet d'atteindre au maximum un score CIDEr de 23,2, ce qui n'est pas significatif dans ce contexte et ne peut être obtenu que par une optimisation abusive de la récompense (*reward hacking*). Des résultats par catégorie de modification sont fournis en annexe C.

5.4 Analyse qualitative

Nous présentons dans la Fig. 1 une analyse comparative des sorties des modèles *IDC* dans différents scénarios : des paires d'images dans la distribution de l'ensemble d'entraînement et de test de EE, et des images hors distribution provenant du web, afin d'évaluer les capacités *zero-shot* des modèles.

Nous observons les limitations de SCORER dans la gestion des images du monde réel dans tous les cas. Il ne parvient à identifier une modification de texte que dans l'échantillon d'entraînement et confond des actions comme l'ajout ou la suppression. CLIP4IDC obtient de bons résultats sur les échantillons de EE mais a du mal avec la généralisation *zero-shot*. Plus précisément, il identifie incorrectement le « set de tennis de table » en raison de son absence dans les données d'entraînement, le confondant avec un objet similaire mais incorrect. De plus, il ne parvient pas à comprendre le

contexte dans lequel l'objet est ajouté. BLIP2IDC, en revanche, excelle sur les données dans la distribution et démontre une capacité supérieure en généralisation grâce à l'utilisation efficace de l'attention croisée. Il reconnaît et décrit avec précision l'ensemble des scènes dans la Fig. 1(c-d), y compris des ajouts complexes comme un set de tennis de table, en tirant parti de son pré-entraînement. Cela permet à BLIP2IDC non seulement d'identifier la présence de feu dans la Fig. 1(d), mais aussi de spécifier l'objet concerné, démontrant ainsi sa capacité à relier les changements sémantiques à leur contexte.

5.5 Limites

BLIP2 est pré-entraîné sur de grands jeux de données image-texte provenant du web, tels que LAION (Schuhmann *et al.*, 2022). Comme BLIP2IDC est une version fine-tunée de BLIP2, il hérite des biais observés dans le modèle BLIP2 en raison de ses données de pré-entraînement.

Notre pipeline d'augmentation synthétique repose également sur plusieurs modèles de génération synthétique et hérite donc également de leurs limitations, telles que la mauvaise représentation ethnique ou les vulnérabilités adversariales. Le LLM utilisé pour générer les variations de vérité terrain peut halluciner, même avec des invites soigneusement formulées.

La tâche elle-même a de nombreuses limites (voir Annexe B). Les modèles d'édition d'images, bien qu'en amélioration, ne sont pas encore capables d'effectuer tous les types de modifications. Une utilisation négligente de ces modèles peut dégrader la qualité du jeu de données si les instructions d'édition ne sont pas bien alignées avec les capacités du modèle. Nous soulignons enfin la limitation liée au domaine : les jeux de données *IDC* ne peuvent guère être exhaustifs dans leur capacité à traiter tous les types de changements, en raison de la multiplicité des interprétations et des descriptions des modifications. Pour un domaine donné, une approche plus adaptée consiste à sélectionner un nombre limité de changements bien définis et à construire à partir de ceux-ci, permettant un meilleur contrôle.

6 Conclusion

Dans cet article, nous proposons un nouveau cadre qui adapte les modèles existants de légendage d'images à la tâche *IDC* afin de bénéficier de leur pré-entraînement étendu. Les données de pré-entraînement pour le légendage d'images sont en effet plus faciles à obtenir que celles pour l'*IDC*. Cela permet de tirer efficacement parti des connaissances globales issues de jeux de données à grande échelle non étiquetés et de les transférer à l'*IDC* en utilisant une quantité plus réduite de données.

À titre de démonstration, nous adaptons BLIP2 à la tâche *IDC*. En encodant de manière innovante les différences entre images au niveau des pixels plutôt qu'en nous appuyant sur le schéma traditionnel à double flux, nous garantissons une approche plus informative et directe pour comprendre les variations entre images. BLIP2IDC permet de surpasser les méthodes existantes sur les benchmarks standards.

De plus, nous introduisons une stratégie d'augmentation synthétique qui non seulement répond aux défis critiques de la rareté des données et du besoin d'architectures robustes applicables au monde réel, mais établit également un nouveau benchmark en matière de performance pour l'*IDC*. Cela ouvre la voie à l'application de l'*IDC* à des données plus diverses et à des types d'éditions plus variés.

Références

BAHDANAU D., CHO K. & BENGIO Y. (2015). Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

BROOKS T., HOLYNSKI A. & EFROS A. A. (2023). Instructpix2pix : Learning to follow image editing instructions. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 18392–18402. DOI : 10.1109/CVPR52729.2023.01764.

CHOUAF S., HOXHA G., SMARA Y. & MELGANI F. (2021). Captioning changes in bi-temporal remote sensing images. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*. DOI : 10.1109/IGARSS47720.2021.9554419.

DOSOVITSKIY A., BEYER L., KOLESNIKOV A., WEISSENBORN D., ZHAI X., UNTERTHINER T., DEHGHANI M., MINDERER M., HEIGOLD G., GELLY S., USZKOREIT J. & HOULSBY N. (2021). An image is worth 16x16 words : Transformers for image recognition at scale. In *International Conference on Learning Representations*.

GOODFELLOW I., POUGET-ABADIE J., MIRZA M., XU B., WARDE-FARLEY D., OZAIR S., COURVILLE A. & BENGIO Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems*.

GUO Z., WANG T.-J. & LAAKSONEN J. (2022). CLIP4IDC : CLIP for image difference captioning. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2 : Short Papers)*.

HOCHREITER S. & SCHMIDHUBER J. (1997). Long Short-Term Memory. *Neural Computation*, p. 1735–1780.

HODOSH M., YOUNG P. & HOCKENMAIER J. (2013). Framing image description as a ranking task : Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*.

HOSSEINZADEH M. & WANG Y. (2021). Image change captioning by learning from an auxiliary task. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 2724–2733. DOI : 10.1109/CVPR46437.2021.00275.

HU E. J., YELONG SHEN, WALLIS P., ALLEN-ZHU Z., LI Y., WANG S., WANG L. & CHEN W. (2022). LoRA : Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

HUANG Q., LIANG Y., WEI J., CAI Y., LIANG H., LEUNG H.-F. & LI Q. (2022). Image difference captioning with instance-level fine-grained feature representation. *IEEE Transactions on Multimedia*. DOI : 10.1109/TMM.2021.3074803.

JHAMTANI H. & BERG-KIRKPATRICK T. (2018). Learning to describe differences between pairs of similar images. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

KIM H., KIM J., LEE H., PARK H. & KIM G. (2021). Viewpoint-agnostic change captioning with cycle consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

KRISHNA R., ZHU Y., GROTH O., JOHNSON J., HATA K., KRAVITZ J., CHEN S., KALANTIDIS Y., LI L.-J., SHAMMA D. A., BERNSTEIN M. S. & FEI-FEI L. (2017). Visual genome : Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, **123**, 32–73. DOI : 10.1007/s11263-016-0981-7.

LAVIE A. & AGARWAL A. (2007). Meteor : an automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation* : Association for Computational Linguistics.

LI J., LI D., SAVARESE S. & HOI S. C. H. (2023). Blip-2 : Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*.

LI J., LI D., XIONG C. & HOI S. C. H. (2022). Blip : Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*.

LI X., YIN X., LI C., ZHANG P., HU X., ZHANG L., WANG L., HU H., DONG L., WEI F., CHOI Y. & GAO J. (2020). Oscar : Object-semantics aligned pre-training for vision-language tasks. In *ECCV*.

LIN C.-Y. (2004). ROUGE : A package for automatic evaluation of summaries. In *Text Summarization Branches Out* : Association for Computational Linguistics.

LIN T.-Y., MAIRE M., BELONGIE S., HAYS J., PERONA P., RAMANAN D., DOLLÁR P. & ZITNICK C. L. (2014). Microsoft coco : Common objects in context. In *Computer Vision – ECCV 2014*.

LIU C., YANG J., QI Z., ZOU Z. & SHI Z. (2023a). Progressive scale-aware network for remote sensing image change captioning. In *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*.

LIU F., YIN C., WU X., GE S., ZHANG P. & SUN X. (2021). Contrastive attention for automatic chest X-ray report generation. In *Findings of the Association for Computational Linguistics : ACL-IJCNLP 2021*.

LIU H., LI C., WU Q. & LEE Y. J. (2023b). Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

MANGRULKAR S., GUGGER S., DEBUT L., BELKADA Y., PAUL S. & BOSSAN B. (2022). Peft : State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft.

MENG C., HE Y., SONG Y., SONG J., WU J., ZHU J.-Y. & ERMON S. (2021). Sdedit : Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*.

MITCHELL M., DODGE J., GOYAL A., YAMAGUCHI K., STRATOS K., HAN X., MENSCH A., BERG A., BERG T. & DAUMÉ III H. (2012). Midge : Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*.

OLIVEIRA DOS SANTOS G., COLOMBINI E. L. & AVILA S. (2021). CIDEr-R : Robust consensus-based image description evaluation. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02.

PARK D. H., DARRELL T. & ROHRBACH A. (2019). Robust change captioning. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, p. 4623–4632.

PARMAR G., SINGH K. K., ZHANG R., LI Y., LU J. & ZHU J.-Y. (2023). Zero-shot image-to-image translation. *ACM SIGGRAPH 2023 Conference Proceedings*.

QIU Y., YAMAMOTO S., NAKASHIMA K., SUZUKI R., IWATA K., KATAOKA H. & SATOH Y. (2021). Describing and localizing multiple changes with transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, p. 1951–1960.

RADFORD A., KIM J. W., HALLACY C., RAMESH A., GOH G., AGARWAL S., SASTRY G., ASKELL A., MISHKIN P., CLARK J., KRUEGER G. & SUTSKEVER I. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

RAMESH A., DHARIWAL P., NICHOL A., CHU C. & CHEN M. (2022). Hierarchical text-conditional image generation with clip latents. *ArXiv*.

ROMBACH R., BLATTMANN A., LORENZ D., ESSER P. & OMMER B. (2021). High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

SCHUHMANN C., BEAUMONT R., VENCU R., GORDON C. W., WIGHTMAN R., CHERTI M., COOMBES T., KATTA A., MULLIS C., WORTSMAN M., SCHRAMOWSKI P., KUNDURTHY S. R., CROWSON K., SCHMIDT L., KACZMARCZYK R. & JITSEV J. (2022). LAION-5b : An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

SHARMA P., DING N., GOODMAN S. & SORICUT R. (2018). Conceptual captions : A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*. DOI : 10.18653/v1/P18-1238.

SHEYNIN S., POLYAK A., SINGER U., KIRSTAIN Y., ZOHAR A., ASHUAL O., PARIKH D. & TAIGMAN Y. (2023). Emu edit : Precise image editing via recognition and generation tasks. *ArXiv*, **abs/2311.10089**.

SHI X., YANG X., GU J., JOTY S. R. & CAI J. (2020a). Finding it at another side : A viewpoint-adapted matching encoder for change captioning. In *European Conference on Computer Vision*.

SHI X., YANG X., GU J., JOTY S. R. & CAI J. (2020b). Finding it at another side : A viewpoint-adapted matching encoder for change captioning. In *European Conference on Computer Vision*.

SUN Y., LI L., YAO T., LU T., ZHENG B., YAN C., ZHANG H., BAO Y., DING G. & SLABAUGH G. (2022). Bidirectional difference locating and semantic consistency reasoning for change captioning. *International Journal of Intelligent Systems*. DOI : https ://doi.org/10.1002/int.22821.

TAN H., DERNONCOURT F., LIN Z., BUI T. & BANSAL M. (2019). Expressing visual relationships via language. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1182.

TU Y., LI L., SU L., DU J., LU K. & HUANG Q. (2023a). Viewpoint-adaptive representation disentanglement network for change captioning. *IEEE Transactions on Image Processing*, **32**, 2620–2635. DOI : 10.1109/TIP.2023.3268004.

TU Y., LI L., SU L., LU K. & HUANG Q. (2023b). Neighborhood contrastive transformer for change captioning. *IEEE Transactions on Multimedia*, **25**, 9518–9529. DOI : 10.1109/TMM.2023.3254162.

TU Y., LI L., SU L., ZHA Z.-J., YAN C. & HUANG Q. (2023c). Self-supervised cross-view representation reconstruction for change captioning. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. DOI : 10.1109/ICCV51070.2023.00263.

TU Y., YAO T., LI L., LOU J., GAO S., YU Z. & YAN C. (2021). Semantic relation-aware difference representation learning for change captioning. In C. ZONG, F. XIA, W. LI & R. NAVIGLI, Éds., *Findings of the Association for Computational Linguistics : ACL-IJCNLP 2021* : Association for Computational Linguistics. DOI : 10.18653/v1/2021.findings-acl.6.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. U. & POLOSUKHIN I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*.

VINYALS O., TOSHEV A., BENGIO S. & ERHAN D. (2015). Show and tell : A neural image caption generator. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

WANG P., YANG A., MEN R., LIN J., BAI S., LI Z., MA J., ZHOU C., ZHOU J. & YANG H. (2022). OFA : Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *Proceedings of the 39th International Conference on Machine Learning*.

WANG W., LV Q., YU W., HONG W., QI J., WANG Y., JI J., YANG Z., ZHAO L., SONG X., XU J., XU B., LI J., DONG Y., DING M. & TANG J. (2024). Cogvlm : Visual expert for pretrained language models.

XU K., BA J., KIROS R., CHO K., COURVILLE A., SALAKHUDINOV R., ZEMEL R. & BENGIO Y. (2015). Show, attend and tell : Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning*, Proceedings of Machine Learning Research : PMLR.

YAO L., WANG W. & JIN Q. (2022). Image difference captioning with pre-training and contrastive learning. In *AAAI Conference on Artificial Intelligence*.

YOUNG P., LAI A., HODOSH M. & HOCKENMAIER J. (2014). From image descriptions to visual denotations : New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*.

ZHANG K., MO L., CHEN W., SUN H. & SU Y. (2023). Magicbrush : A manually annotated dataset for instruction-guided image editing. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS)*.

ZHOU K., YANG J., LOY C. C. & LIU Z. (2022). Learning to prompt for vision-language models. *International Journal of Computer Vision*.

A Jeux de données existants

CLEVR-Change. Introduit par *Robust Change Captioning* (Park *et al.*, 2019), l'objectif de ce jeu de données est d'évaluer les capacités fondamentales de compréhension visuelle, telles que la capacité à identifier la forme, la couleur, le matériau et la position d'un objet, ainsi que de déterminer si l'une ou aucune de ces propriétés a changé. Ce jeu de données permet d'évaluer si un modèle *IDC* comprend les relations spatiales et s'il est robuste aux changements non sémantiques. Cinq types de modifications de scène sont définis. Les instructions d'édition et les légendes de référence sont automatiquement construites selon un modèle. Cependant, ce jeu de données a été créé avec un moteur 3D, et tous les éléments sont des scènes 3D éloignées des conditions du monde réel, ce qui conduit à un modèle qui ne se transfère pas très bien aux exemples concrets.

CLEVR-DC. CLEVR-DC (Kim *et al.*, 2021) est un dérivé de CLEVR-Change avec un changement extrême de points de vue. Il utilise les mêmes types de modifications que CLEVR-Change. L'objectif de ce jeu de données est d'évaluer l'invariance aux changements de points de vue du modèle *IDC*, mais il partage les limitations du jeu de données original.

Spot-The-Diff (STD). Contrairement à CLEVR-Change, STD (Jhamtani & Berg-Kirkpatrick, 2018) utilise un cadre réel pour les paires d'images. Pour faire face au problème de production d'images modifiées, il utilise l'évolution temporelle. Ainsi, une paire d'images est composée d'une image prise à un instant donné et d'une autre image du même lieu prise à un instant ultérieur, avec un point de vue fixe. La plupart des images sont situées dans un parking, ce qui restreint fortement la diversité des vérités terrain, car la plupart des changements dans un parking consistent à ajouter ou supprimer des personnes ou des voitures. Bien que fortement biaisé, ce jeu de données permet de tester les modèles *IDC* sur des images du monde réel.

Image Editing Request (IER). Né de l'exploration de Reddit et Zhopped, le jeu de données IER (Tan *et al.*, 2019) repose sur des instructions d'utilisateurs en ligne dans des forums spécialisés Photoshop. L'image d'entrée et l'instruction d'édition étaient publiées, et les utilisateurs envoyaient leurs images modifiées sur Reddit. L'exploration de ces résultats d'édition manuelle a conduit à un jeu de données de très haute qualité et diversifié, mais avec une échelle très réduite. Chaque paire d'images de l'ensemble de test possède trois légendes de référence, écrites par trois annotateurs différents. IER offre une large gamme de modifications sémantiques, comme par exemple « ajouter un chapeau de marin aux canards », « remplacer l'arrière-plan par un vaisseau spatial » ou « remplacer les balais par des sabres laser ».

MagicBrush. Ce jeu de données est le premier jeu de données d'édition d'images guidée par des instructions, annoté manuellement et à grande échelle (Zhang *et al.*, 2023), couvrant divers scénarios d'édition globale et locale. MagicBrush comprend 10 000 triplets $(I_{ref}, T_{instruction}, I_{edited})$, ce qui est suffisant pour fine-tuner des modèles d'édition d'images à grande échelle avec l'utilisation de LoRA. Pour ce jeu de données, les modifications ne sont pas regroupées en catégories comme dans CLEVR-Change, ce qui rend difficile l'explication des performances et le ciblage de faiblesses spécifiques.

Emu Edit (EE). Ce jeu de données (Sheynin *et al.*, 2023) de Meta est composé de paires d'images originales issues de MSCOCO (Lin *et al.*, 2014) et d'images modifiées générées par le modèle Emu Edit, un modèle de génération performant, qui prend une image et une instruction d'édition en entrée. Ce jeu de données contient également des informations détaillées sur le type de modifications générées, regroupées en huit catégories différentes : Ajout, Suppression,Texte, Couleur, Arrière-plan,

Style, modification Locale, et Globale. Associé au cadre réel des images originales, qui permet une plus grande utilisabilité, ces propriétés en font un excellent candidat pour un benchmark *IDC* exigeant.

B Problèmes liés à la tâche

Les modèles *IDC* sont entraînés sur des triplets $(I_{ref}, I_{modified}, GT)$ et évalués sur un ensemble de test de paires $(I_{ref}, I_{modified})$ en comparant la légende générée avec les légendes de référence GT, à l'aide de diverses métriques d'évaluation automatiques, telles que BLEU, ROUGE, METEOR et CIDEr (Papineni *et al.*, 2002; Lin, 2004; Lavie & Agarwal, 2007; Oliveira dos Santos *et al.*, 2021). Certains aspects des jeux de données du monde réel affaiblissent la qualité de l'évaluation.

Métriques sous-optimales. Les métriques basées sur les références évaluent plus précisément la qualité des phrases générées si plusieurs références sont fournies. Cinq références est généralement le minimum utilisé. Cependant, aucun des jeux de données réalistes disponibles ne fournit un nombre cohérent et suffisant de références, rendant les évaluations moins précises.

Manque de cohérence dans les vérités terrain. Dans STD et IER, les légendes de référence annotées manuellement ne font pas systématiquement référence à la même différence ou au même ensemble de différences, en particulier lorsque plusieurs modifications sont apportées à l'image. Certaines légendes décrivent une seule modification, d'autres plusieurs. Un modèle *IDC* entraîné pour décrire uniquement la différence principale aura du mal avec ce type de vérité terrain, et c'est le cas pour tous les modèles *IDC* existants. Soit toutes les différences doivent être mentionnées dans chaque vérité terrain, soit une seule, de manière cohérente, pour éviter des objectifs contradictoires.

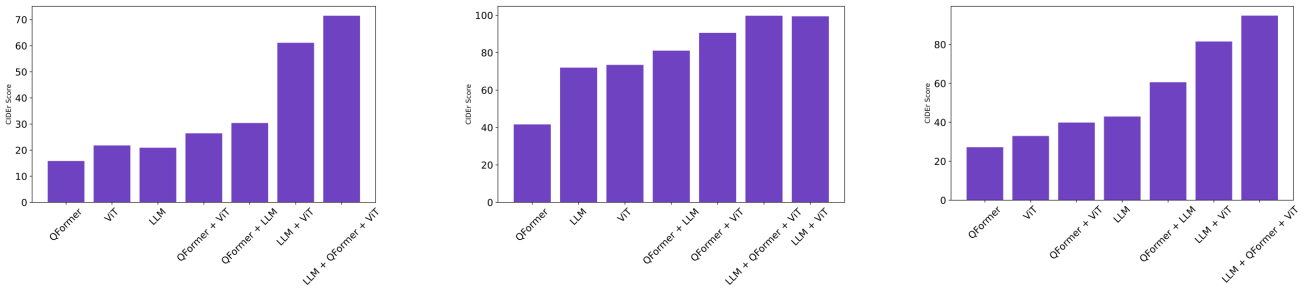
Jeux de données du monde réel de petite taille. Étant donné que les jeux de données du monde réel sont très chronophages à produire, ils sont de petite taille, avec au maximum 10 000 paires texte-image. En conséquence, ils peuvent conduire à des modèles spécialisés dans des sous-ensembles spécifiques de modifications possibles, comme pour STD. À l'inverse, lorsque les modifications sont très diverses, comme pour IER, un type de modification peut ne survenir que quelques fois. Les métriques dépendent alors fortement des divisions de données, selon qu'une modification apparaît uniquement dans l'ensemble d'entraînement ou dans l'ensemble de test. Cela est d'autant plus possible si les modifications ne sont pas regroupées en classes permettant un échantillonnage stratifié.

Catégorisation des modifications. Regrouper les changements en catégories est utile pour l'échantillonnage stratifié, mais aussi pour analyser les résultats, car cela facilite l'identification des changements difficiles et des faiblesses des modèles *IDC*.

C Résultats supplémentaires

C.1 Étude d'ablation

Les variations de performance (score CIDEr) entre les différentes configurations de BLIP2IDC, en fonction des modules fine-tunés, sont illustrées dans la Fig. 5. Cette analyse met en évidence l'impact significatif du fine-tuning des modules ViT et LLM. Étant donné le décalage entre les modules ViT et LLM pour les domaines d'entrée et de sortie, il n'est pas surprenant que le QFormer se



(a) IER score (b) EE Score (c) Syned Score

FIGURE 5 – Ablation de BLIP2IDC, mesurée par CIDEr, selon les datasets de fine-tuning des modules.

distingue comme le composant le moins critique pour le fine-tuning, car il a déjà subi un entraînement approfondi lors de la phase de pré-entraînement de BLIP2.

C.2 Reśultats par catégorie de modification

TABLE 5 – Scores CIDEr par catégorie de changement et version du jeu de données. Ce tableau présente le score CIDEr pour chacune des catégories suivantes : Ajout (A), Texte (T), Arrière-plan (B), Couleur (C), Style (S), Global (G), Suppression (R), Local (L) et Intégral (I), ce dernier étant le CIDEr calculé sur l'ensemble des échantillons de l'ensemble de test.

Dataset	A	T	B	C	S	G	R	L	I
Syned	102.45	147.17	111.78	151.87	12.77	**31.56**	72.27	61.00	94.83
EE	101.01	**147.69**	**119.48**	155.76	30.53	29.46	93.39	76.24	100.83
Syned+ EE	**107.38**	147.42	112.37	**164.40**	**37.07**	30.43	**111.16**	**77.81**	**106.83**

Dans la Table 5, nous étudions le comportement de BLIP2IDC en fonction de ses données d'entraînement, qu'il s'agisse de Syned, EE ou des deux, pour différentes catégories de modifications. La combinaison des jeux de données est généralement la meilleure configuration globale, à l'exception de la catégorie Texte (T). Nous supposons que cela est dû à la faible capacité de la plupart des modèles d'édition d'images à générer du texte dans les images, ce qui entrave les performances de l'augmentation synthétique. Les catégories les plus faibles en termes de performance du score CIDEr sont les catégories Style et Global, qui sont des changements très subjectifs.