Enhancing Text-to-Video Editing with Motion Map Injection

Seong-Hun Jeong*[1], In-Hwan Jin*[1], Haesoo Choo*[1], Hyeonjun Na*[1] and Kyeongbo Kong†[2]
[1]Pukyong National University, [2]Pusan National University
tlqwkrk915@pukyong.ac.kr, {bds06081, tndi28, wewe1st}@naver.com, kbkong@pusan.ac.kr

Abstract

Based on the remarkable performance of text-to-image diffusion models, text-guided video editing studies recently have been expanded. Existing video editing studies have introduced an implicit method of adding cross-frame attention to estimate inter-frame attention, resulting in temporally consistent videos. However, because these methods use models pre-trained on text-image pair data, they do not handle unique property of video: motion. When editing a video with prompts, the attention map of the prompt implying the motion of the video (e.g. 'running', 'moving') is prone to be poorly estimated, which causes inaccurate video editing. To address this problem, we propose the 'Motion Map Injection' (MMI) module to consider motion explicitly. The MMI module provides text-to-video (T2V) models a simple but effective way to convey motion in three steps: 1) extracting motion map, 2) calculating the similarity between the motion map and the attention map of each prompt, and 3) injecting motion map into the attention maps. Given experimental results, input video can be edited accurately with MMI module. To the best of our knowledge, our study is the first method that utilizes the motion in video for text-to-video editing. Extensive experimental results are in https://currycurry915.github.io/MMI/.

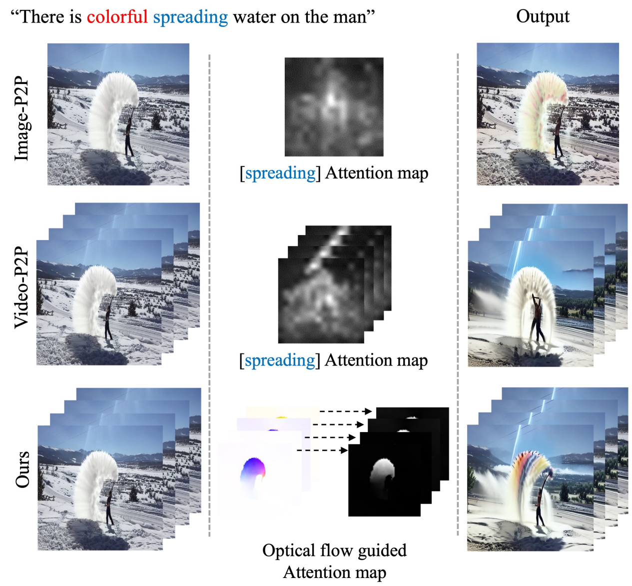

Figure 1: Comparison of attention map and edited video output derived from each existing method. Image-P2P [1] and Video-P2P [2] failed to estimate the attention map, resulting in discrepancy between prompt and video. Our method edited video realistically by exploiting optical flow guided attention maps.

1. Introduction

Recent research on text guided diffusion models and large-scale language models has led to unprecedented advances in image generation and image editing. In the field of image editing, various studies [1, 3, 4, 5, 6, 7, 8, 9, 10, 11] have been conducted. Among them, Prompt-to-Prompt (P2P) [1] offers users intuitive image editing by proposing several methods and applications to control the attention map that identifies the semantic relationship between prompt token and input image.

Research on diffusion model has been expanded to text-guided video editing tasks. Most of these studies [2, 12, 13, 14] use text-guided image diffusion models for baseline. However, since a video consists of several frames unlike images, it should be edited in consideration of temporal information. To this end, existing video editing studies [14, 15, 16] have devised an implicit method of adding cross-frame attention to estimate inter-frame attention, resulting in temporal consistent editing. Recently, starting with Video-P2P [2] vid2vid-zero [16], and FateZero [17], research on image based P2P [1] have been expanded to video editing.

A vast amount of text-video pair dataset is required to train the implicit structures that use cross-frame attention to understand temporal information. However, there are limit of data to train them. That is why the most of video editing algorithms have chosen the fine-tuning method of image diffusion model [12] for the input video or zero-shot approach [14] to edit video in consideration of temporal information. Because the image diffusion model was trained

*equal contribution
†Corresponding author

on the text-image pair dataset, it is not good at estimating the attention map for prompts containing motion as shown in Fig 1. As a result, it is difficult to perform appropriate attention control for the prompt, which reduces the capability of editing videos. One solution is to estimate motion information explicitly.

Optical flow is the information of pixels that have moved between frames of video (i.e. 'motion'). There are a lot of optical flow estimation methods that have been used in wide range of video tasks. With existing state-of-the-art optical flow estimation network [18], we can obtain highly accurate motion which can be injected into video editing model.

In this paper, we propose a framework to complement attention maps with optical flow to make video editing more effective and accurate. It consists of the following three steps. Firstly, existing optical flow estimation algorithm [18] estimates motion from the input video. Secondly, with the motion, attention maps of off-the-shelf video diffusion model are complemented. Lastly, modify the video by applying the improved attention map.

The most important part of this framework is first and second steps. We propose 'Motion Map Injection (MMI)' module for these steps. Specifically, MMI module has three steps. Firstly, module estimates motion map which is magnitude of the optical flow. Secondly, the similarities between the motion map and attention maps of each prompt are computed with template matching algorithm [19]. Finally, the motion map is multiplied by the similarity weights, and added to the every attention map of each prompt. However, the image diffusion model does not accurately estimate the attention map of the motion prompt, so the attention map of motion prompt such as 'moving' is directly replaced with the motion map without template matching. Our proposed MMI module is a efficient method that can supplement motion information in the prompt without learning a vast video dataset, which can dramatically improve the editing of the existing video editing framework.

2. Proposed Method

Let be input video $\mathcal{V}$ which consists of n frames. We define the source prompt as $\mathcal{P}$ just like the Prompt-to-Prompt [1] setting. In $\mathcal{P}$, the prompt containing motion information of the video is called motion prompt $\mathcal{P}_{\mathcal{M}}$ (e.g. 'running', 'moving').

In this section, we provide an overview of our framework, which is illustrated in Fig. 2(a). The attention map $\mathcal{A}_{\mathcal{P}}$ representing the correlation between the frame of $\mathcal{V}$ and $\mathcal{P}$ is estimated through the cross-attention layer in the T2V-Model. Our proposed MMI module estimates the optical flow V_{flow} of $\mathcal{V}$ using the pre-trained optical flow estimation network [18] in the first step. In the second step, a motion map is obtained by applying L2 norm to the optical flow. After calculating the correlation between the attention map of all prompts and the motion map using NCC, it complements the existing attention map according to the correlation. Lastly, the motion map is inserted into the Attention Map of the motion prompt $\mathcal{P}_{\mathcal{M}}$. Using this approach, the MMI module provides the optical flow information of the video to the T2V-Model, enabling improved video editing capabilities.

2.1. Preliminary

Prompt-to-Prompt P2P [1] based on text-guided diffusion model edits image by modifying source prompt $\mathcal{P}$. Cross-attention layer inside the diffusion model produces cross-attention map indicating spatial correlation between visual and textual features. Spatial features of noise image $\phi(z_t)$ are projected on query matrix $Q = \ell_Q(\phi(z_t))$ and key matrix $K = \ell_K(\psi(\mathcal{P}))$ through learned linear projections ℓ_Q and ℓ_K. Attention map A can be written as

$$A = Softmax\left(\frac{QK^T}{\sqrt{d}}\right), \tag{1}$$

where d denotes latent projection dimension of K and Q. Because P2P [1] can edit images by controlling these attentions, it gives intuitive image editing by modifying only $\mathcal{P}$. As in Fig. 2(b), P2P [1] can edit images by replacing A of the word to be edited in $\mathcal{P}$ with the attention of the target prompt or by providing an additional A.

2.2. Motion Map Injection Module

Motion Map Extraction In this paper, we use pre-trained optical flow estimation algorithm, UniMatch [18], to estimate motion of video. Firstly, we calculate the correlation for the pixels of the two frames by matrix product and then normalize for the last two dimensions using the softmax function. Then, matching distribution D_{flow} is obtained for each pixel location in F_{t-1} with respect to all pixel locations in F_t

$$D_{\text{flow}}(i,j,k,l) = \mathcal{S}\left(\frac{\sum_{d'=1}^{d} F_{t-1}(i,j,d') \cdot F_t(k,l,d')}{\sqrt{N}}\right), \tag{2}$$

where N denotes normalization coefficient to prevent the value from increasing after internal operation, and $\mathcal{S}$ represents the softmax function. We calculate the weighted average of the matching distribution D_{flow} on the 2D coordinates of the pixel grid G_{2D} to obtain correspondence $\hat{G}_{2D}$. Finally, the optical flow V_{flow} can be obtained by subtracting the two grids. We obtain the motion map V^*_{flow} by applying the L2 norm to the optical flow V_{flow} .

Motion Map Injection using NCC It is necessary to calculate the similarity of each other to effectively supplement the existing attention map using the estimated motion map. To this end, we use the template matching algorithm [20]. We set the source image to attention map and the target im-

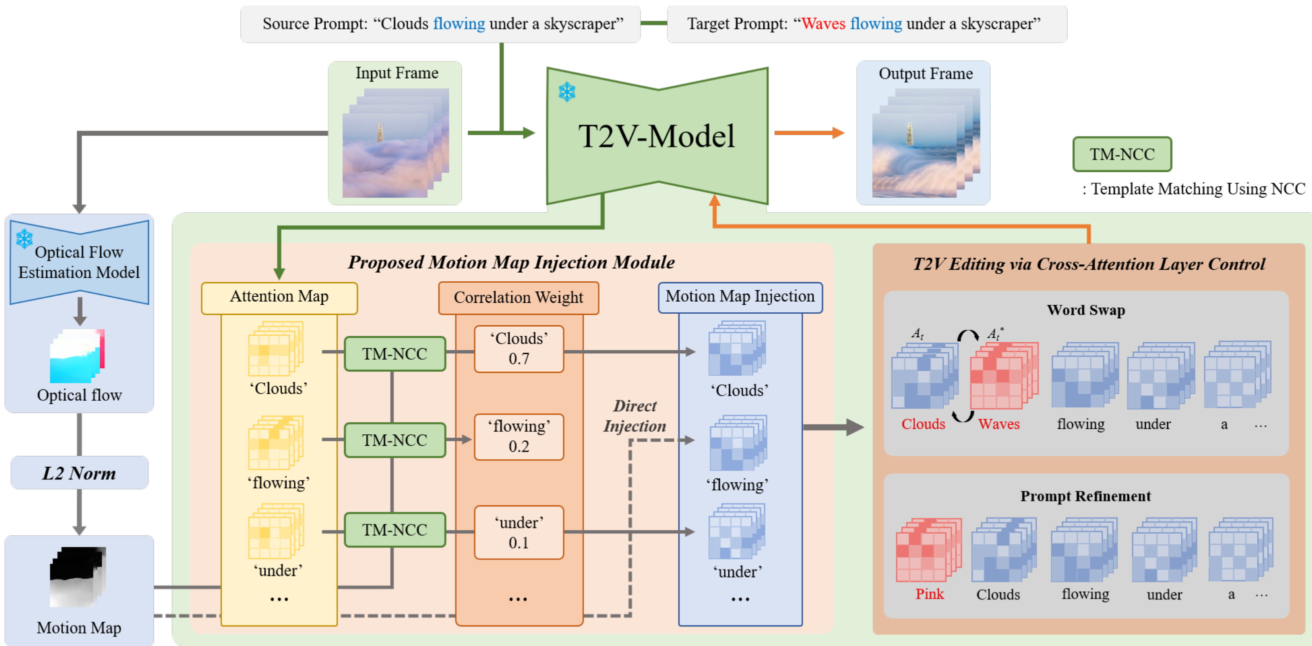

Figure 2: Overall framework of this study. First, the T2V-Model generates an attention map by receiving video and prompts as input. At the same time, the Motion Map Injection module receives the video frame, generates a motion map, and injects it into the attention map of the T2V-Model. After that, text-to-video editing is performed using the attention map that includes video motion information.

age to motion map and calculate the Normalized Cross Correlation (NCC) as below

$$C_k = \frac{1}{n}\sum_{x,y}\frac{(V^*_{flow}(x,y) - \bar{V}^*_{flow})(A_k(x,y) - \bar{A}_k)}{\sigma_V \sigma_A}, \quad (3)$$

where n denotes the number of pixels, x, y denotes pixel of source image and target image, $\bar{V}^*_{flow}$ and $\bar{A}_k$ denote the average of pixels of V^*_{flow} and A_k, σ_V and σ_A denote the standard deviation of pixels of V^*_{flow} and A_k. k denotes the index of a particular word in the entire prompt. We calculate the correlation score by performing template matching at every denoising step t of the diffusion model, making the sizes of V^*_{flow} and A_k alike for all k.

Finally, as shown in (4), a new attention map is identified by injecting $C_k \cdot V^*_{flow}$ to attention map of each prompt

$$A^*_k = A_k + \lambda \cdot \frac{(C_k \cdot V^*_{flow})}{t}, \quad (4)$$

where λ is a hyperparameter for motion map injection rate, t denotes denoising step. Here, the image diffusion model does not accurately estimate the attention map of the motion prompt token such as 'flowing', so the motion map is directly inserted instead of template matching.

3. Experiments

Baseline model We tested our module to the Video-P2P [2], vid2vid-zero [16], and FateZero [17], which can intuitively edits a video with text by manipulating a cross attention map representing the relationship between text and video.

Table 1: Generally, when MMI module was applied, the metrics were improved.

	CLIP Score	Masked PSNR	BRISQUE
Video-p2p	28.50	23.33	37.21
+ Ours	**30.15**	**24.92**	**32.11**
vid2vid-zero	29.60	19.81	15.99
+ Ours	**31.90**	**20.16**	**15.09**
FateZero	26.60	**27.72**	32.16
+ Ours	**28.45**	27.30	**29.60**

Evaluation Metrics CLIP Score [21] and masked PSNR [2] were used to evaluate textual similarity and region preservation. We assessed the quality of the edited videos with BRISQUE [22]. Details of these metrics are at Section 5.1 of supplement.

3.1. Quantitative Results

The quantitative results of this study are in Table 1.We compared 20 video data and 3 baseline models with 3 metrics. CLIP Score [21] was measured by comparing target prompt with edited image. The results confirmed that our method was better at understanding meaning of prompt compared to the other models. For masked PSNR[2], the performance of our framework scored slightly higher in

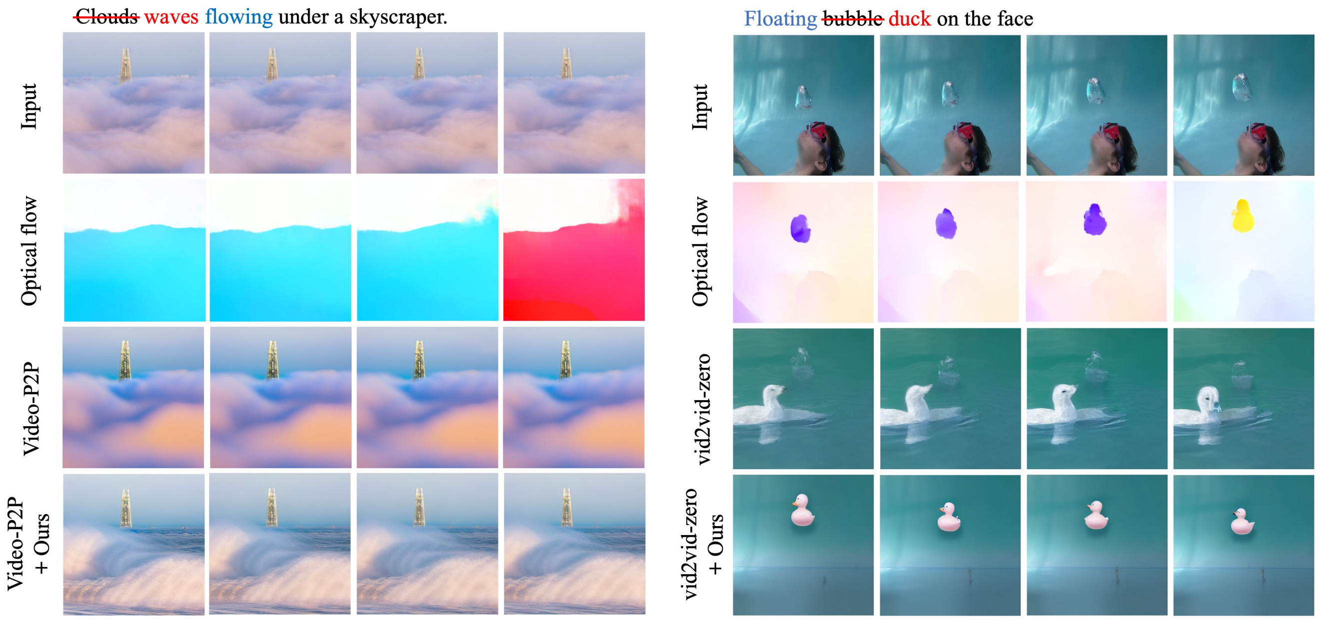

Figure 3: Qualitative results of our study. This shows the result of two examples corresponding to each different model [2, 16]. The top row is the input video frame, the second row is the optical flow extracted through the optical flow estimation model [18], the third row is the output of [2] and [16], and the last row is the result of applying our module. The model utilizing the motion information of the video perform better editing. Additional results are at our project page.

Video-P2P [2] and vid2vid-zero [16]. This means that the existing models edited more in areas not intended than our proposed model. Using one of the No-Reference Image Quality Assessments, BRISQUE [22], Our model was lower scores than others. It indicates our module improves the quality of edited video.

3.2. Qualitative Results

Figure 3 shows the qualitative results of the Video-P2P [2], the vid2vid-zero [16], and those utilizing proposed MMI. Previous methods tend to produce structurally unnatural images, but our method produces structurally coherent and content-preserving frames. In left example, due to inaccurate estimation of the attention map for the motion prompt, it was unable to perform editing that aligns with the target prompt. However, with the injection of the motion map using our module, we achieved better performance in editing. In right one, the result is poorer compared to Video-P2P [2]. Nevertheless, our model enhances attention maps and enables control them via target prompts. There are more results on the project page attached to the abstract.

3.3. User Study

To compare our proposed model with 3 existing models, we conducted user study on three criteria: structure preserving, text alignment, and realism & quality. We presented input video, target prompt, and videos edited with each model to users. Between two edited videos, the outputs were evaluated by making users choose more faithful videos for the presented criteria. The results filled in Table 2 show that the output of our module is more preferable. The supplement section 6 contains a detailed protocol for the user study.

Table 2: User study for (a) Video-P2P, (b) vid2vid-zero, and (c) FateZero, and our framework on three criteria. The units of each number mean %.

User Preference (%)	(a)	(b)	(c)	Ours
Text Alignment	6.11	4.00	6.22	**83.67**
Structure Preserving	9.38	9.66	10.33	**70.63**
Realism & Quality	9.50	9.22	10.33	**70.95**

4. Conclusion

We propose a Motion Map Injection (MMI) module to inject estimated motion map into the attention map of the image diffusion model. Injecting motion map into attention map makes general video editing performance improved because attention map of motion prompt was enhanced. This was proven by improved metrics and showed by qualitative results.

5. Acknowledgements

This research was supported in part by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2020R1A6A3A01098940, 2021R1I1A1A01051225).

References

[1] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 2

[2] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. *arXiv preprint arXiv:2303.04761*, 2023. 1, 3, 4

[3] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 1

[4] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023. 1

[5] Chen Henry Wu and Fernando De la Torre. Making text-to-image diffusion models zero-shot image-to-image editors by inferring" random seeds". In *NeurIPS 2022 Workshop on Score-Based Methods*. 1

[6] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, SIGGRAPH '23. Association for Computing Machinery, 2023. 1

[7] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022. 1

[8] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. In *The Eleventh International Conference on Learning Representations*, 2023. 1

[9] Jooyoung Choi, Yunjey Choi, Yunji Kim, Junho Kim, and Sungroh Yoon. Custom-edit: Text-guided image editing with customized diffusion models. *arXiv preprint arXiv:2305.15779*, 2023. 1

[10] Zhixing Zhang, Ligong Han, Arnab Ghosh, Dimitris N Metaxas, and Jian Ren. Sine: Single image editing with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6027–6037, 2023. 1

[11] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 1

[12] Eyal Molad, Eliahu Horwitz, Dani Valevski, Alex Rav Acha, Yossi Matias, Yael Pritch, Yaniv Leviathan, and Yedid Hoshen. Dreamix: Video diffusion models are general video editors. *arXiv preprint arXiv:2302.01329*, 2023. 1

[13] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Weixian Lei, Yuchao Gu, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. *arXiv preprint arXiv:2212.11565*, 2022. 1

[14] Wen Wang, Kangyang Xie, Zide Liu, Hao Chen, Yue Cao, Xinlong Wang, and Chunhua Shen. Zero-shot video editing using off-the-shelf image diffusion models. *arXiv preprint arXiv:2303.17599*, 2023. 1

[15] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. *arXiv preprint arXiv:2306.07954*, 2023. 1

[16] Wen Wang, kangyang Xie, Zide Liu, Hao Chen, Yue Cao, Xinlong Wang, and Chunhua Shen. Zero-shot video editing using off-the-shelf image diffusion models. *arXiv preprint arXiv:2303.17599*, 2023. 1, 3, 4

[17] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. *arXiv preprint arXiv:2303.09535*, 2023. 1, 3

[18] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2, 4

[19] Artiom Basulto-Lantsova, José Alfredo Padilla-Medina, Francisco Javier Pérez-Pinal, and Alejandro Israel Barranco Gutiérrez. Performance comparative of opencv template matching method on jetson tx2 and jetson nano developer kits. *2020 10th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 0812–0816, 2020. 2

[20] John P Lewis. Fast template matching. In *Vision interface*, volume 95, pages 15–19. Quebec City, QC, Canada, 1995. 2

[21] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528. Association for Computational Linguistics, nov 2021. 3

[22] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012. 3, 4