Supplementary material
JAWS: Just A Wild Shot for Cinematic Transfer in Neural Radiance Fields

Xi Wang[1]*, Robin Courant[1,2]*, Jinglei Shi[3], Eric Marchand[1], Marc Christie[1]
[1]Inria, IRISA, CNRS, Univ. Rennes, [2]LIX, Ecole Polytechnique, IP Paris, [3]VCIP, CS, Nankai Univ.
xi.wang@inria.fr, robin.courant@polytechnique.edu, jinglei.shi@nankai.edu.cn
{eric.marchand, marc.christie}@irisa.fr

Prelude: First, we would like to thank all reviewers for the efforts in reviewing our paper. In this document, additional materials are provided to supplement our main paper. We provide the ethical and societal discussion in Sec. A. Then in Sec. C.2, besides the photometric L2 loss investigated in the main paper, we study more types of on-screen metrics which are often used to train deep networks for similar tasks, including image Structural Similarity (SSIM) and perceptual VGG loss, showing limitations of each metric and the advantage of our proposed on-screen loss. In Sec. C.3, we take a further step from on-screen loss to cinematic loss, demonstrating how the on-screen term and inter-frame term in cinematic losses work together to realize the cinematic transfer task by showing ablation studies. Finally, in Sec. C.4, we show a pipeline about how to use our system to assist graphics engine animation shooting workflow. **We have also attached a detailed demo video to this document, which illustrates more animated materials on qualitative and ablation experiments, therefore we strongly recommend the reviewers to watch this companion material.**

A. Ethical and societal discussion

Societal impact. JAWS globally makes 3D animation shooting more accessible, by granting users an off-the-shelf camera motion transfer tool (see Sec. 5.1 of the main manuscript). Our system benefit both inexperienced users, in terms of enabling the accessibility to create cinematic movie clips without knowing expertise knowledge, and professional artists on alleviating the repetitive works and accelerating their design workflow.

On the other hand, we are aware that such approaches could be harmful for the creativity in general. More specifically, as well as new state-of-the-art generative methods [1, 2], a drawback of JAWS could be the limiting of human spontaneous creativity, by relying only on numerical tools and prejudged paradigms from reference films. In addition, copyrights and manipulated content concerns can rapidly arise.

Environmental impact. By considering that we used pre-trained NeRF, pose estimator and flow estimator models during this project, the main computational cost is due to the inference and backpropagation to the camera parameters. We ran our method for experiments (demos) and ablations study. Consider (i) that we have run 20 demos once, and 5 different ablations with 5 times each; (ii) we limit optimization iterations to 200 for a pair of frames; and (iii) an iteration takes 3 seconds. In total, this project has led to $3 * 200 * (20 + (5 * 5)) = 27,000$ GPU seconds, i.e. 450 hours. With one NVIDIA RTXA5000, it results in 230 W $*$ 450 h $= 103.5$ kWh, and then 58 gCO2/kWH $*$ 103.5 kWh $= 6,003$ gCO2 $= 6$ kgCO2 emitted.

B. Implementation details

We ran all our experiments on a NVIDIA RTXA500, the experiment is done with our implementation of instant-NGP [3] derived from Pytorch [4] and 'torch-ngp' [5]. We use 8 Layers for the HashEncoder with $fp16$ mixed precision to economize memory usage, with learning rate of 0.014. For more implementation details, we will release the code and relevant dataset (*e.g.* Unity scene).

All the '*working images*' (image utilized during optimization) are synthesized with resolution of 224 height pixel and the width varies according to the reference aspect ratio, 256 sampling steps along one ray for accelerating the computation. For

*Equal contribution. Corresponding to jinglei.shi@nankai.edu.cn.

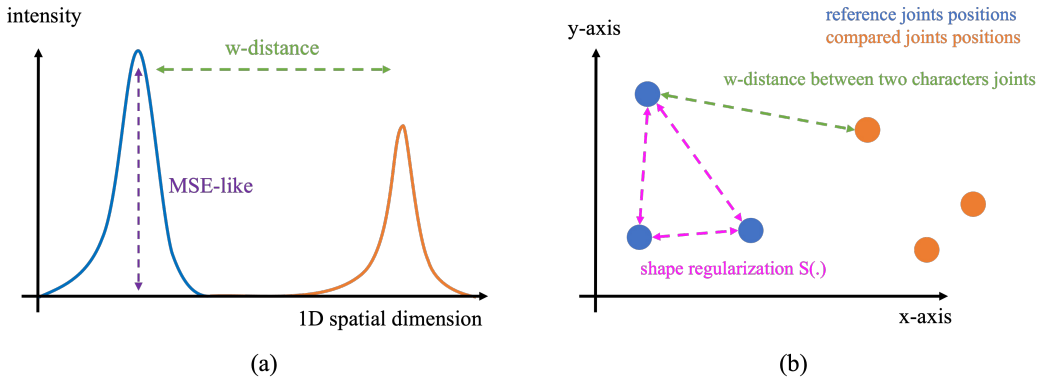

Figure 1. (a) a brief 1-dimensional signal demonstration of the difference between the MSE-like loss and the w-distance loss which focus more on spatial distance between two signals analogical to differentiable Euclidean distance; (b) an exemplary illustration of Eq. 1, the different functionality of the term $d_w(\mathbf{H}_i^*, \hat{\mathbf{H}}_i)$ (green) and the shape regularization $S(\mathbf{H})$ (purple) on two on-screen joints distributions (orange and blue).

the '*demo images*' (image utilized for illustration purposes), we use 'nerfstudio' [6] to render high quality frames based on pre-computed camera poses.

C. Additional experimental results

C.1. Heatmap loss of Cinematic loss

To compute a differentiable distance between two heatmaps, we use Wasserstein distance [7] (w-distance for short). The choice of w-distance aims to highlight the on-screen spatial relation (similar to on-image Euclidean joints distance) rather than MSE-like measures emphasizing on intensity difference (see more in [8]). By minimizing the w-distance between two character pose heatmaps, we are able to transfer the framing information from a target reference to a differentiable rendering space. Demonstration in Fig. 1. subfig. (a) also depicts the difference between a 1D MSE-like intensity-oriented loss and a 1D w-distance spatial-oriented loss.

Lets consider a confidence heatmap $\mathbf{H} \in (\mathbb{R}^+)^{H \times W \times J}$ with J being joint number on human skeleton (see Fig. **??**) generated from a pose estimation network, where each channel represents a detection probability of a given joint in the image domain. To calculate the character pose loss $\mathcal{L}_{\text{pose}}$, we compute w-distance d_w between confidence maps of a reference $\mathbf{H}^*$ and a synthetic view $\hat{\mathbf{H}}$ on each channel.

$$\begin{aligned} \mathcal{L}_{\text{pose}} &= \sum_{i=1}^{J} d_w(\mathbf{H}_i^*, \hat{\mathbf{H}}_i) + ||S(\mathbf{H}^*) - S(\hat{\mathbf{H}})|| \\ \text{where} \quad S(\mathbf{H}) &= \sum_{i=1}^{J} \sum_{j=1}^{J} d_w(\mathbf{H}_i, \mathbf{H}_j) \end{aligned} \tag{1}$$

The regularization S is defined as an inter-joint matrix measured in w-distance of each joint $\mathbf{H_j}$ to all the others from the *same* heatmap $\mathbf{H}$. This term assures the inter-joint shape similarity between heatmaps while optimizing the framing. See the Fig. 1, subfig. (b), it demonstrates the different function of two terms, one aims to approach in-total distance between two poses (green) joint-by-joint, and another assures the shape similarity of each pose by minimizing the distance of inter-joint distance matrix (similar to a covariance matrix).

C.2. On-screen loss comparison

We show in our main paper (Fig. III) the distribution of normalized error when using photometric L2 loss (i.e. MSE) and our pose loss. In this section, we compare more losses which are widely used to train deep networks for relevant tasks, showing the robustness of our pose loss under the cinematic context. The extra two losses selected here are image Structural Similarity (SSIM loss) [9] and human perceptual loss using VGG encoder (VGG loss) [10]. SSIM measures the similarity between two images in terms of image contour and texture, while the VGG loss focus more about the perceptual quality from human beings; it compares the intermediate feature maps within a pretrained encoder (VGG encoder in our

case) between generated image and reference image. Note that both two metrics measure not solely the pixel intensity, but also the global quality and abstract gist of image information, hence they are representative losses often used to help various computer vision tasks such as image style transfer, super-resolution etc. As a complement to Fig. III in the main paper, we illustrate in Fig. 2 the loss distribution when perturbing the camera pose around a reference position on x and y translational directions $\Delta x \in [-0.3, 0.3], \Delta y \in [-0.3, 0.3]$, within same or different scene against the reference. Only our pose loss gives meaningful convergence cones on both scenes, which proves the performance across different scenes and the ability of anchoring the character on the screen under different scenarios.

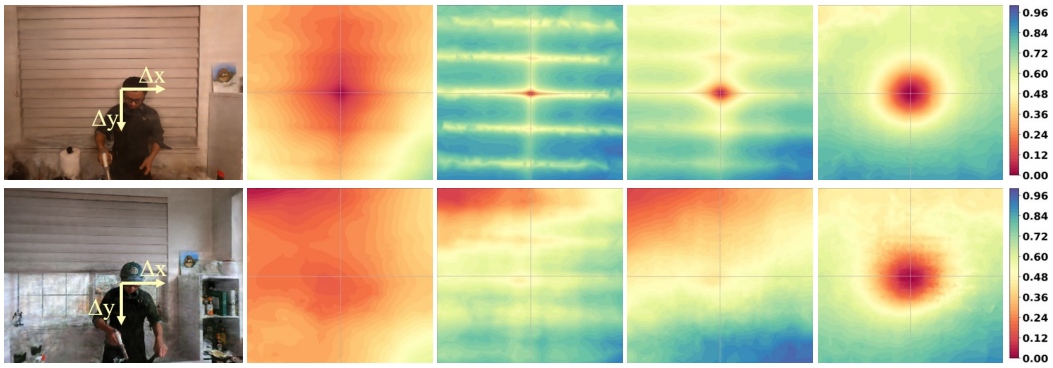

Figure 2. Normalized error distribution when using different metrics. We show in the 1st row the errors on the *same* scene, and the 2nd row the results on the *different* scene to the reference (reference is generated from the scene of the 1st row).

In the main paper, we presented the ablation study results of our proposed methods by comparing them with iNeRF (pixel photometric loss). In this section, we provide some qualitative experimental results based on VGG-based perception loss. Due to numerical instability, the convergence cone is unable to guide optimization towards the correct gradient direction, resulting in numerical explosions (yielding directly to *NaN*). Figure 3 illustrates the qualitative ablation study results obtained when using VGG-based perception loss. As shown, the numerical instability leads to misleading optimized camera positions, deviating from the definition of NeRF, and causing *NaN* numerical explosions.

C.3. Ablation study on cinematic loss

In the previous section, we showed that our proposed pose loss can anchor well the characters information, ensuring a low level of on-scene composition even under different environments. In this section, we analyze the results of ablation of losses

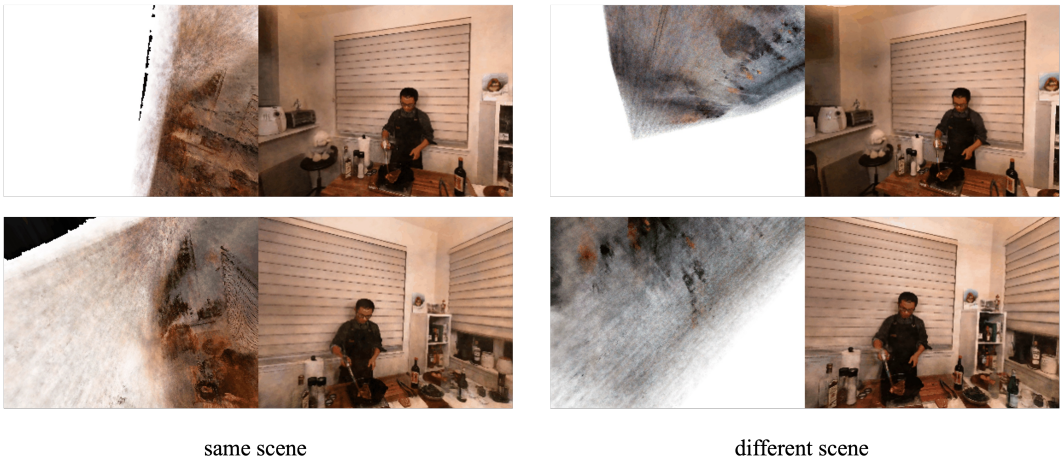

Figure 3. Qualitative ablation study results obtained when using VGG-based perception loss under same and different reference scenes. Numerical instability leads to misleading optimized camera positions, deviating from the definition of NeRF and causing *NaN* numerical explosions.

and explain the observations to show the effectiveness of combining the on-screen pose loss with the inter-frame flow loss. This section can be considered as a complementary part for Sec.6 and Tab.1 of our main paper. The purpose of the ablation is to transfer the motion from a NeRF generated reference to different scenes (same and different to the reference one) under light and strong perturbations (*i.e*. close or far initialization positions).

Fig. 4, Fig. 5. Fig. 6 and Fig. 7 illustrate the synthesized frames when using different types of losses, along with corresponding optical flows. It is strongly recommended to zoom in figures for better visibility. Observations and analysis are as follows:

- Both iNeRF [11] (losses on only sampled pixel) and Pixel loss (using guidance map to deactivate gradient of the sampled pixel but computing loss on whole image) can produce views similar to the reference sequence by simply minimizing photometric differences.
- Pose loss alone anchors the character in the scene wrt the character's position, but it fails to capture the inter-frame information (see optical flows), the synthesized frames suffer from a 'jerky' effect (which can be seen by the incorrect optical flows between frames and shaky video effect, please do view the demo for better animated results), as it merely concentrates on the pose information instead of the relative motion between other frames.
- Although flow loss alone can well mimic the relative movement of the camera (inter-frame information) by showing good results when the initialization position is close, due to neglecting the on-screen information it suffer bias (*i.e*. drifts which fails to minimize the error between the target and reference) when starting from a far intialization to the correct one.
- Finally, our proposed cinematic loss combines the advantages of the flow loss (on relative tracking) and pose loss (on anchor composition), it simultaneously captures both inter-frame and on-screen information and can produce sequences with similar characteristics to the reference clip.

We now evaluate the capacity of our approach to transfer cinematic motion to *different* contents. Fig. 5, Fig. 6 and Fig. 7 show the results of this ablation. It is noteworthy that both iNeRF loss and Pixel loss approaches which perform well when reapplied to the same scene, now totally fail under different scene contents. They either render frames prone to artifacts, or even collapse during the inference by giving numerical errors (see Fig. 5). The reason is that photometric loss is no longer a reliable metric when applied to a distinct content. Shortcomings of only pose loss (inter-frame information lost) or only flow loss (character deviation) still remain. On the contrary, our proposed combined loss keeps robust performance on the different scene contents in terms of both inter-frame (optical flow results) and on-screen information (composition framing).

Though limits are shown in the ablation as well, as displayed in Fig. 7, the leaning posture of the character seems to influence the camera roll angle at the end of the sequence (see last row). This shows how the algorithm tries to transfer simultaneously the flow and the pose information. But due to the strong difference on the poses between two scenes, the final performance are comprised as worse than the results of the other scenes yet still far better than pixel-based methods.

C.4. Additional qualitative results

In Fig. 8, we show the detailed structure of the '*movies to 3D engine*' application presented in Sec. 5 of the main manuscript. We start from a standard 3D scene in Unity, train a NeRF model by Unity rendered multi-view images, and apply JAWS to generate a camera trajectory mimicking the visual effects of an in-the-wild reference clip.

Once the trajectory is computed, it can be re-implemented into Unity. Different to many end-to-end style generation tasks, our generated trajectory can be used directly for animation shooting or be easily modified as a intuitive and compatible warm-start basis of the whole workflow. For additional qualitative results, example of this application, please watch the attached demo video.

C.5. Optimization time

We report in the following table the **optimization time** (in seconds per optimization step) wrt the number of samples in the guidance map, which is tested on a 224×298 image. Usually one keyframe requires 50 to 200 steps depending on the desired quality. For example, for ~ 250 frames (interpolated from 20 keyframes) and 100 steps, our method takes ~ 50 mins.

no. sample	2,000	4,000	8,000	16,000	32,000	65,000
sec per step	1.59	1.58	1.77	1.97	2.30	2.22

References

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 1

[2] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1

[3] T. Müller, A. Evans, C. Schied, and A. Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 2022. 1

[4] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Adv. Neural Inform. Process. Syst.*, 2019. 1

[5] J. Tang. Torch-ngp: a pytorch implementation of instant-ngp, 2022. https://github.com/ashawkey/torch-ngp. 1

[6] M. Tancik*, E. Weber*, E. Ng*, R. Li, B. Yi, T. Wang, A. Kristoffersen, J. Austin, K. Salahi, A. Ahuja, D. McAllister, and A. Kanazawa. Nerfstudio: A framework for neural radiance field development, 2022. 2

[7] L. V. Kantorovich. Mathematical methods of organizing and planning production. *Management science*, 1960. 2

[8] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *Int. conf. on Machine Learning*, 2017. 2

[9] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, 2004. 2

[10] J. Justin, A. Alexandre, and F. Li. Perceptual losses for real-time style transfer and super-resolution. In *Eur. Conf. Comput. Vis.*, pages 694–711, 2016. 2

[11] L. Yen-Chen, P. Florence, J. T. Barron, A. Rodriguez, P. Isola, and T. Lin. inerf: Inverting neural radiance fields for pose estimation. In *IEEE Int. Conf. on Intelligent Robots and Systems*, 2021. 4

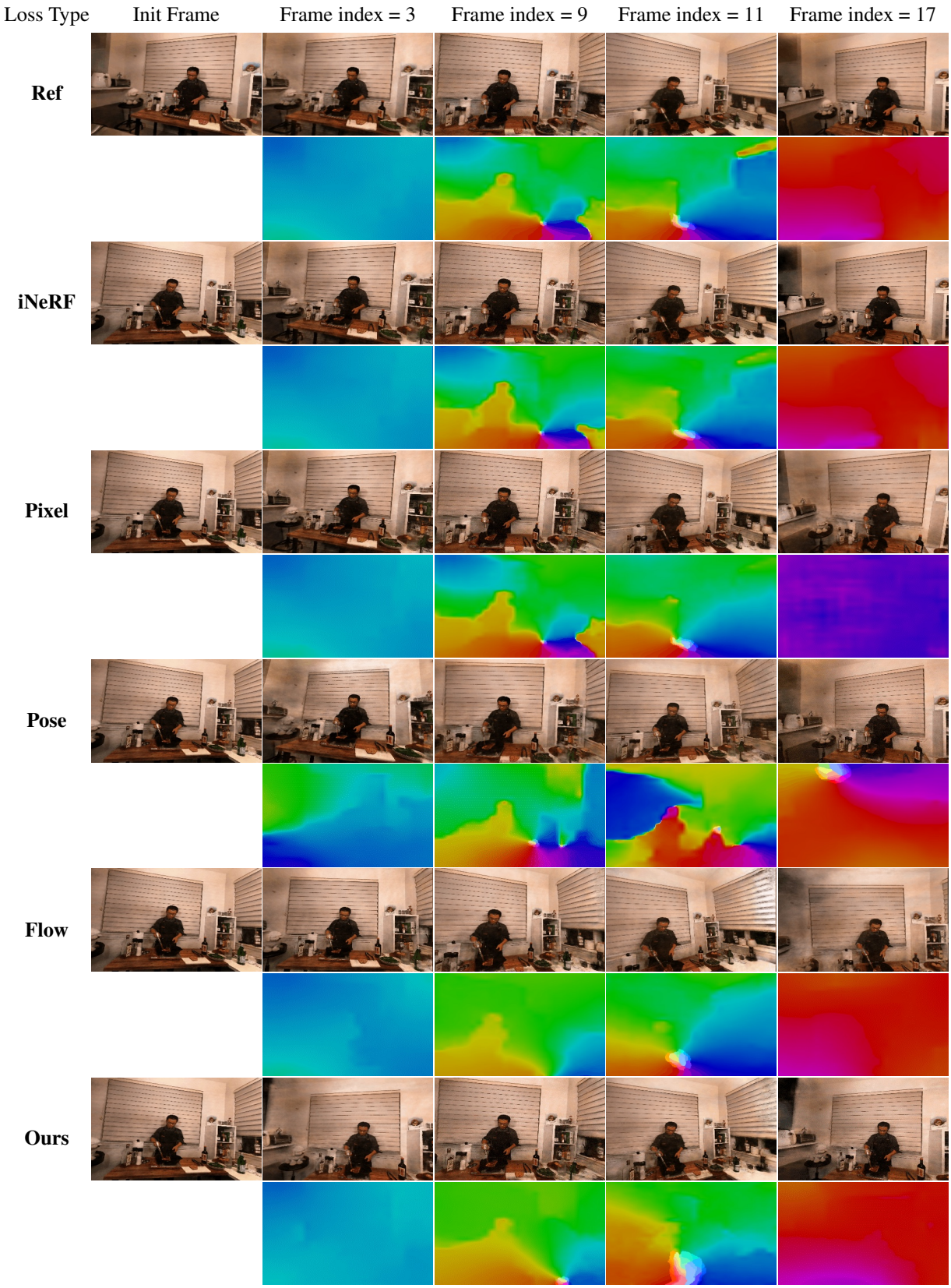

Figure 4. Scene *Steak Flame*; comparing the capacity of different techniques to transfer the cinematic features of the reference clip (first row) to another (similar) scene.

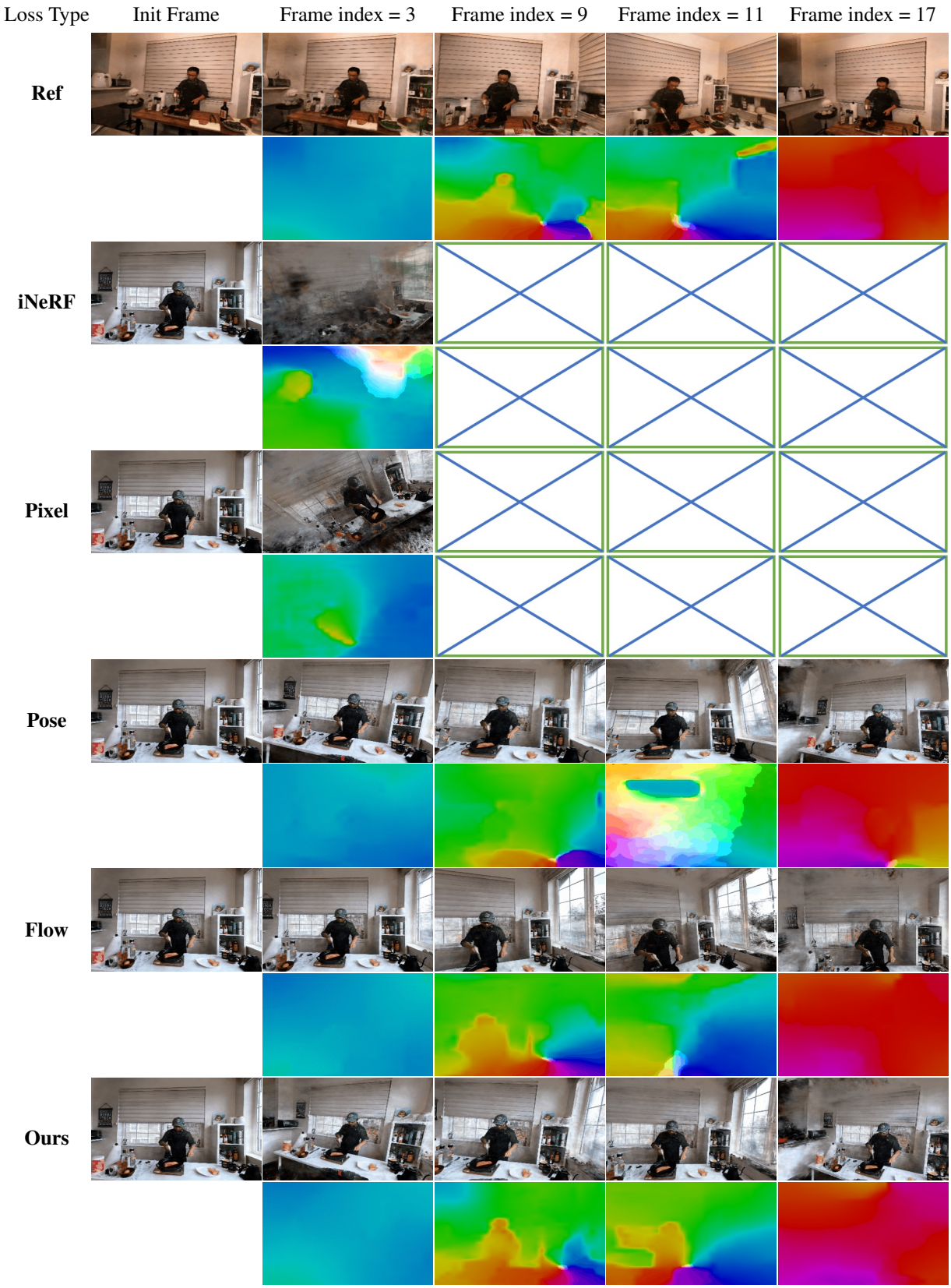

Figure 5. Scene *Salmon Flame*; comparing the capacity of different techniques to transfer the cinematic features of the reference clip (first row) to another (similar) scene.

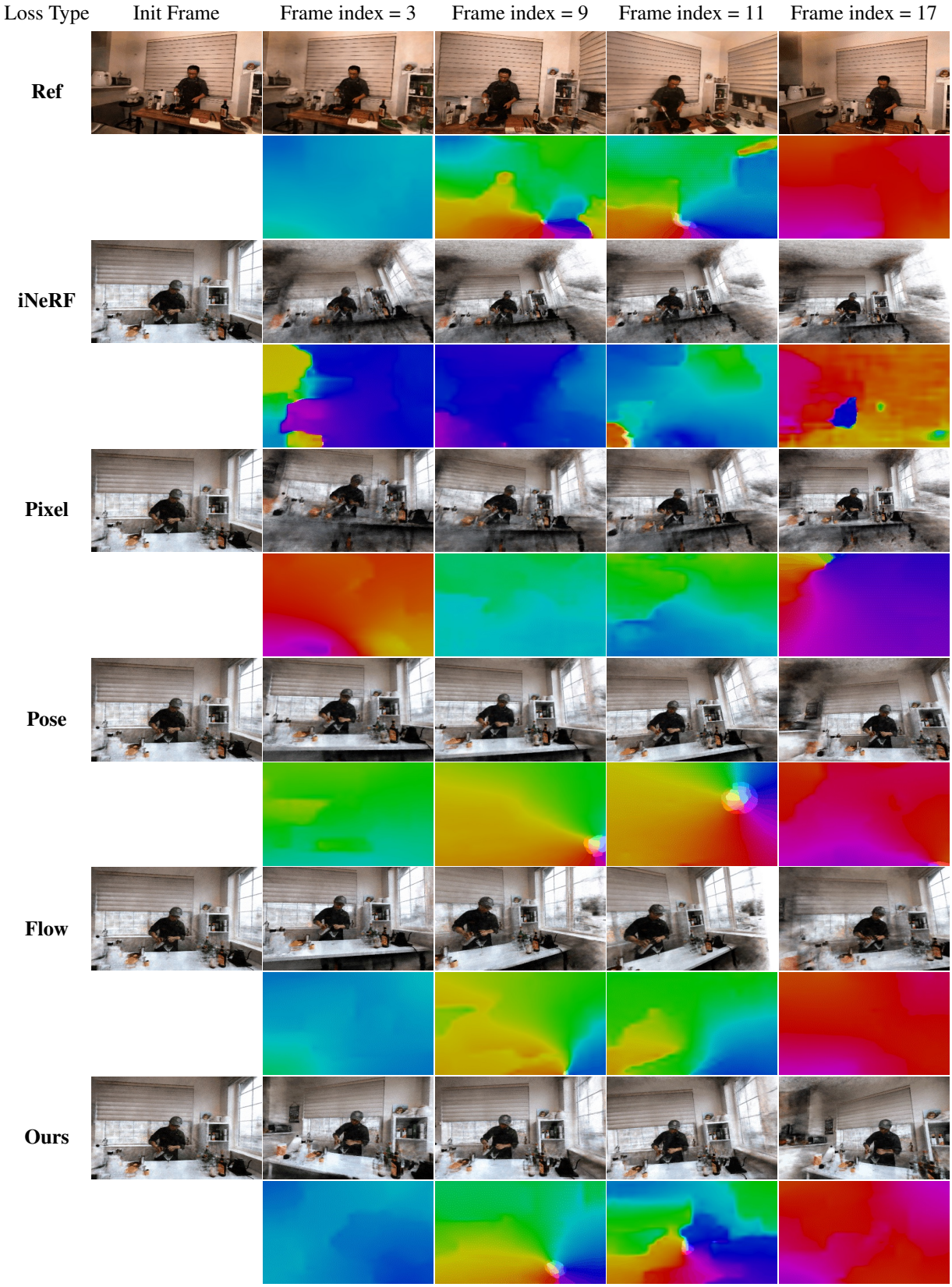

Figure 6. Scene *Coffee Martini*; comparing the capacity of different techniques to transfer the cinematic features of the reference clip (first row) to another (similar) scene.

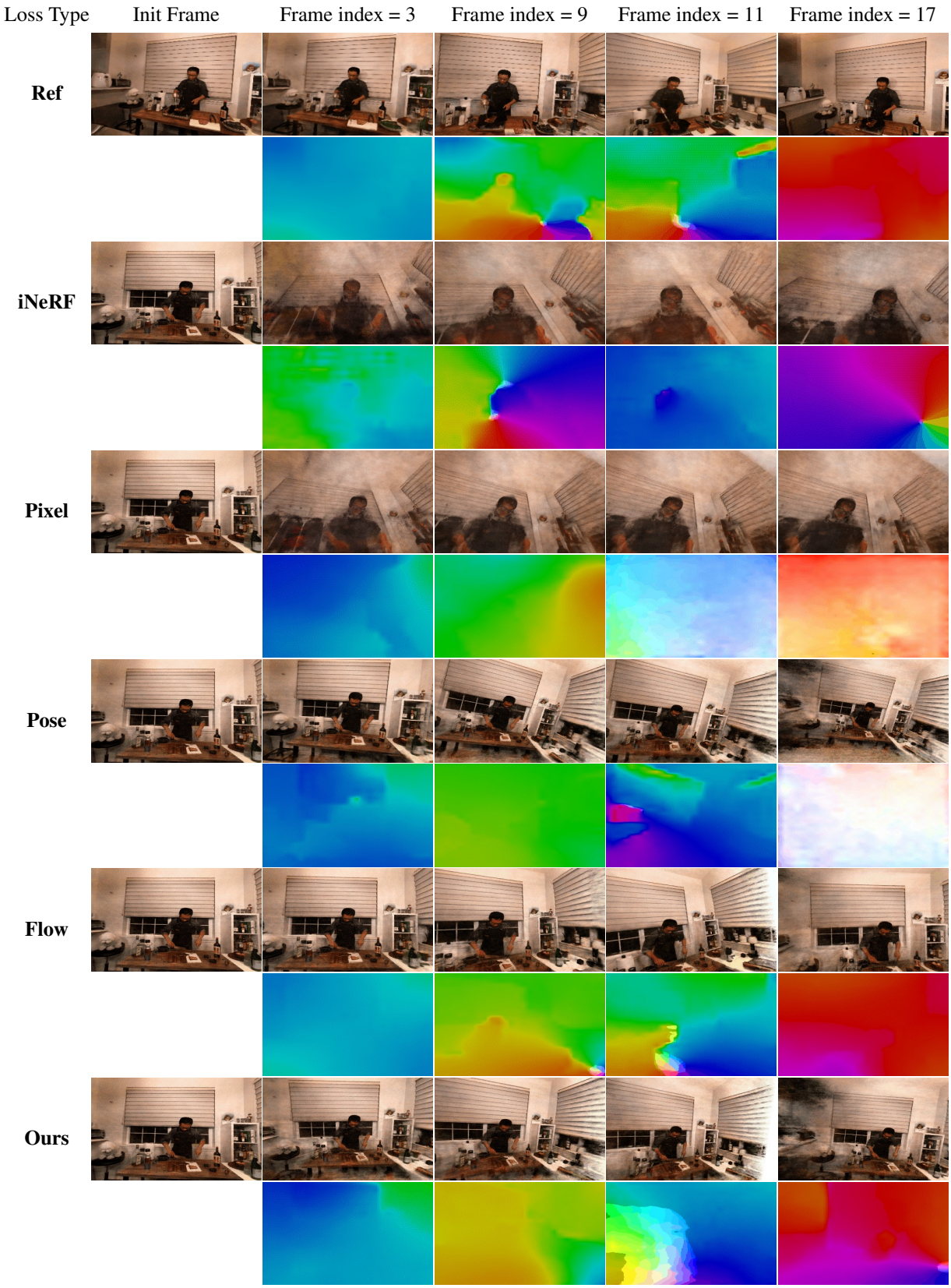

Figure 7. Scene *Cut Roasted Beef*: comparing the capacity of different techniques to transfer the cinematic features of the reference clip (first row) to another (similar) scene.

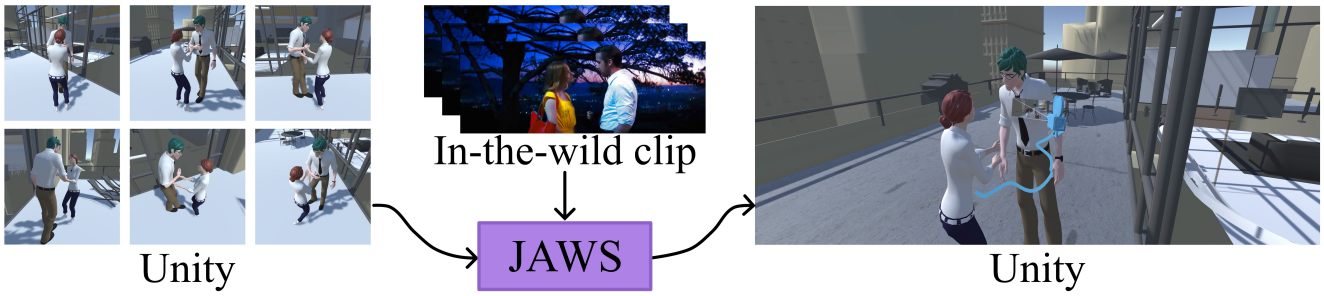

Figure 8. Overview of the *movies to 3D engine* pipeline described in Sec. 5 of the main manuscript.