Eurographics Symposium on Rendering 2025
B. Wang and A. Wilkie
(Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 44 (2025), Number 4

StructuReiser: A Structure-preserving Video Stylization Method (Supplementary Material)

R. Spetlik[1], D. Futschik[2], D. Sýkora[1]

[1]Czech Technical University in Prague, Faculty of Electrical Engineering, Czech Republic
[2]Google Research, USA

In this supplementary material, we first present additional results and comparisons with the current state-of-the-art in text-driven as well as keyframe-based video stylization (Sec. S1) and present more results on the real-time video call stylization scenario where we demonstrate differences between varying strength of adherence to the structural elements (Sec. S2). Sec. S3 shows an alternative 3-axis plot of perceptual study, and in Sec. S4 a threshold analysis of 75 % just-noticeable difference is presented.

S1. Additional results and comparisons

In this section, we present further comparisons of our method with state-of-the-art video stylization approaches. The results of additional sequences that compare our method with text-driven approaches are presented in Figures S1–S6 and sequences comparing our method with keyframe-based techniques in Figures S7–S8. In all cases, the additional results appear to align with the discussion in the main paper (Sec. 4), i.e., our approach tends to preserve more structural details than the compared text-driven and keyframe-based methods (see also our supplementary video).

In Fig. S9, we present a comparison between our approach and the method of Kim et al. [KLC*24]. In the case of the sequence shown in the left half of the figure, our approach appears to be better aligned with the structure of the target frame $\mathbf{y}$. For the sequence in the right half of the figure, the result is not as distinctive, therefore we ask the reader to see our supplementary video. As seen in the supplementary video, our method exhibits fewer temporal artifacts and maintains consistent facial features across challenging frames, indicating higher temporal stability under these test conditions.

We would like to clarify the difference between the conditioning used in our method and the one used by Kim et al. [KLC*24]. In our method, we do not rely on text guidance. Instead, in each training iteration, we inject the target-frame structure by sampling from the diffusion model's prior at a specific denoising step t. This ensures that the final stylized frame $\hat{\mathbf{y}}$ adheres closely to the structure of the target $\mathbf{y}$. In contrast, the work of Kim et al. [KLC*24] is built on the method of Brooks et al. [BHE23] that uses two-fold conditioning: the input image and the text instruction. Each conditioning has a guidance scale with which the degree of similarity between the generated samples and the input image is balanced, as well as the degree of similarity with the editing instruction. Both conditions are used during the entire image generation process. The method of Kim et al. [KLC*24] has a significant memory footprint. Therefore, only 90 stylized frames are available for each sequence featured in our supplementary video.

S2. Real-time video call scenario

Texler et al. [TFK*20] proposed a real-time video call scenario in which the appearance of the call participant is stylized in real time. In this scenario, only a single artist-made keyframe is used to train our method and the methods of Texler et al. [TFK*20] and Futschik et al. [FKL*21]. In Fig. S10, we observe a key advantage of our method: By increasing the strength of the loss term $\mathcal{L}_{\text{structure}}$, the structure of the target frame $\mathbf{y}$ is prioritized, and by decreasing its strength, the style characteristics of the style exemplar are emphasized. Compared to the results of Texler et al. [TFK*20] (a) and Futschik et al. [FKL*21] (b), we find that our results preserve structural elements present in the target frame $\mathbf{y}$ more consistently. To emphasize this fact, we provide three different settings: (c) emphasize more style features of the stylized keyframe, (d) achieve a balance between style and structure, and (e) prioritize structural details present in the target frame.

S3. Perceptual study – alternative plot

Here, we present an alternative plot depicting the results of perceptual study from the main paper. Instead of colored heatmap, there is a three-axis plot in Fig. S11.

S4. Threshold Analysis of 75 % Just-Noticeable Difference

In a two-alternative forced-choice (2-AFC) test a viewer sees two clips and must pick the one that looks *better*. If the two clips are objectively identical, viewers will guess, giving the correct answer half the time (50 %). As the quality gap between the two clips grows, the probability of picking the better one rises. The **75 % JND** is the point where observers pick the correct clip three times out of four. At that gap the difference is *just noticeable* for most people.

© 2025 The Author(s). Computer Graphics Forum published by Eurographics - The European Association for Computer Graphics and John Wiley & Sons Ltd.
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

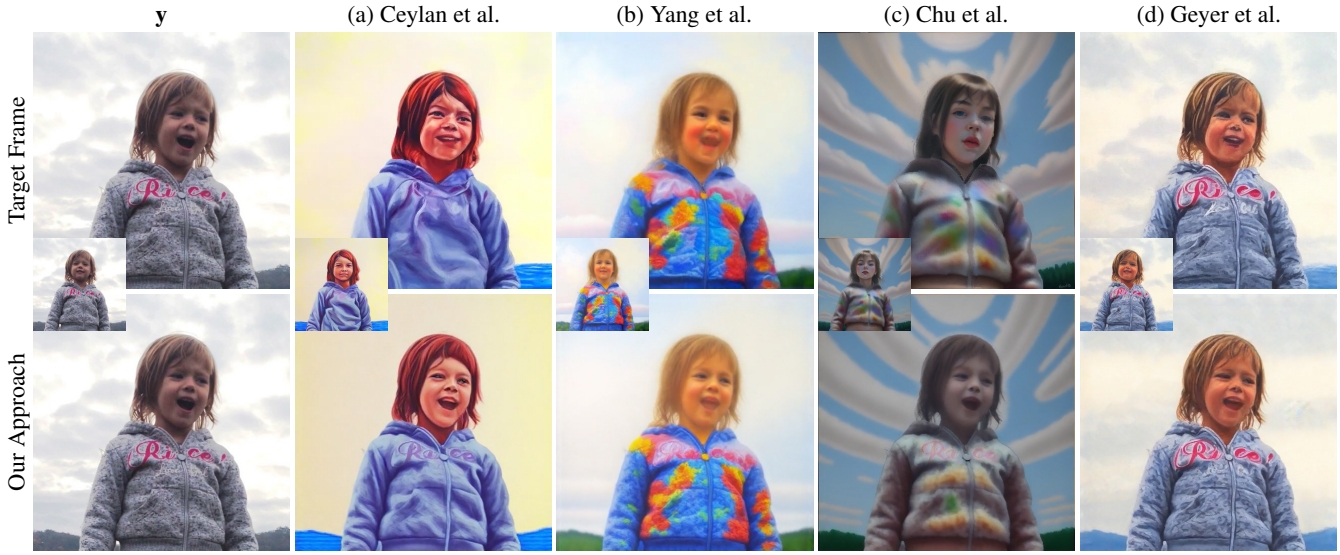


Figure S1: *Comparison with the state-of-the-art in text-driven video stylization: The target video sequence (see a representative target frame* $\mathbf{y}$*) has been stylized using text-driven approaches (top row): (a) Ceylan et al. [CHM23], (b) Yang et al. [YZLL23], (c) Chu et al. [CHLC24], and (d) Geyer et al. [GBTBD24]. One frame from those stylized sequences was used as a keyframe (see small insets). The style of this keyframe has been propagated to the rest of the target sequence* $\mathbf{y} \in \mathcal{Y}$ *using our approach (bottom row). Note how our approach better preserves structural details seen in the target frame. Also, see our supplementary video to compare consistency across the entire sequence. As illustrated in the supplementary video, text-driven approaches often exhibit visible frame-to-frame flicker in certain regions. By contrast, our method reduces flicker, resulting in more consistent structure.*

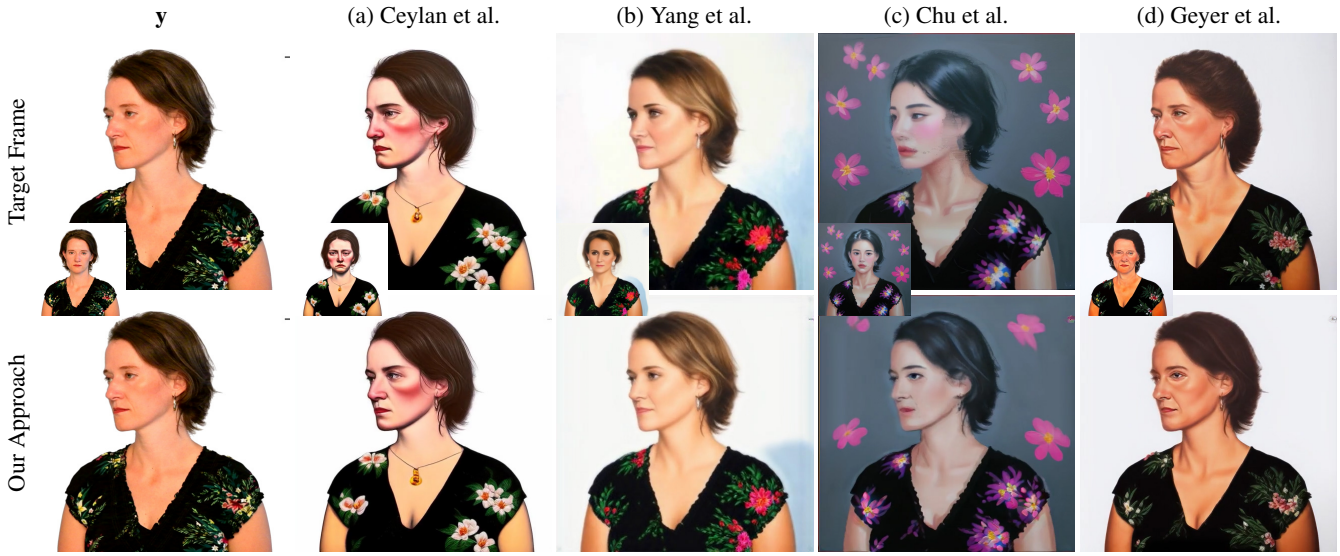


Figure S2: *Comparison with the state-of-the-art in text-driven video stylization (cont.): See Fig. S1 for a detailed explanation.*

© 2025 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.

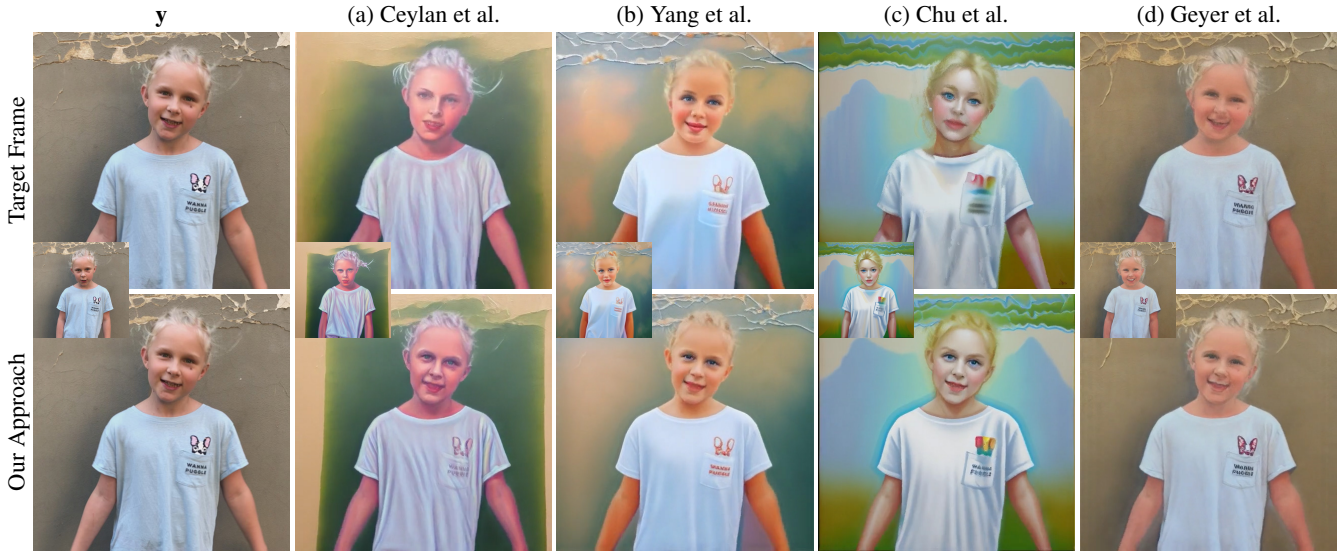


Figure S3: *Comparison with the state-of-the-art in text-driven video stylization (cont.): See Fig. S1 for a detailed explanation.*

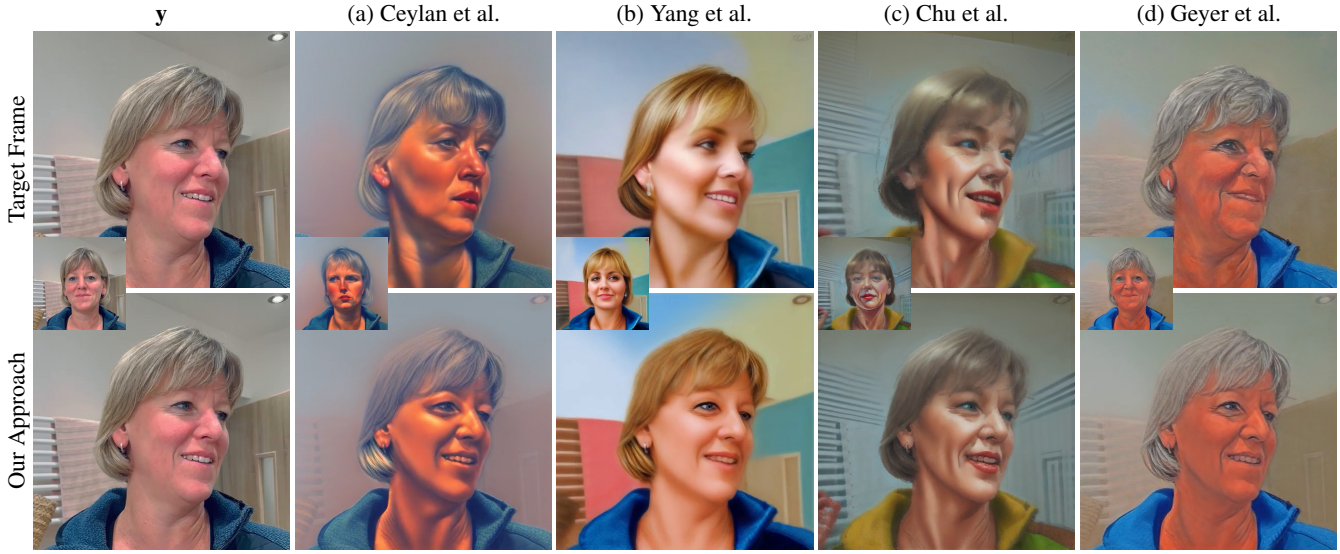


Figure S4: *Comparison with the state-of-the-art in text-driven video stylization (cont.): See Fig. S1 for detailed explanation.*

© 2025 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.

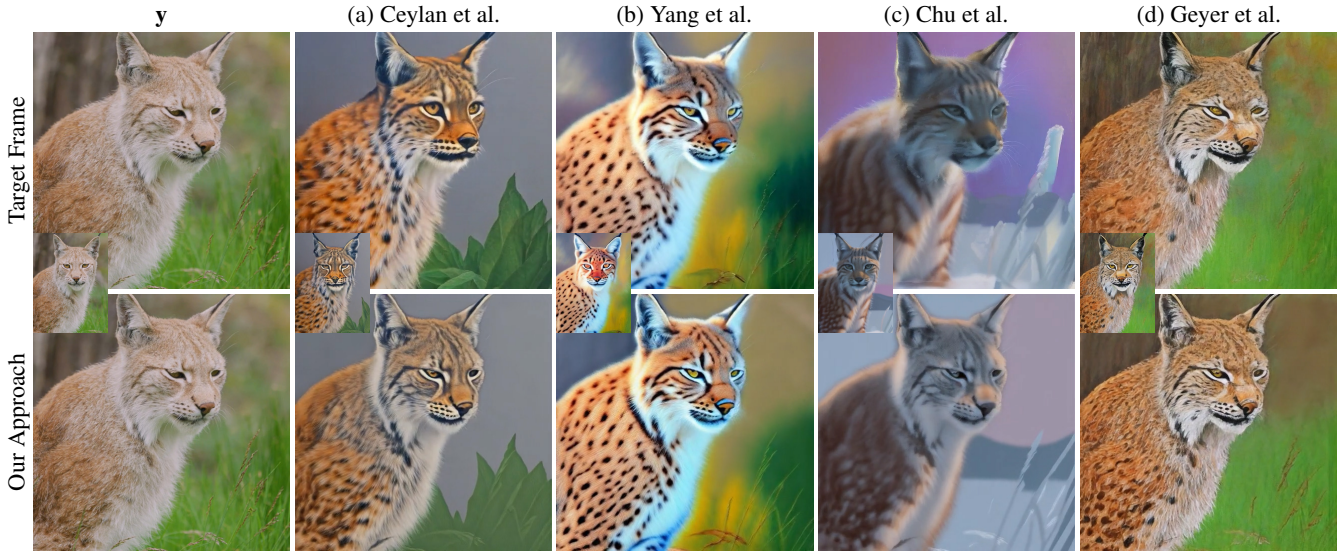
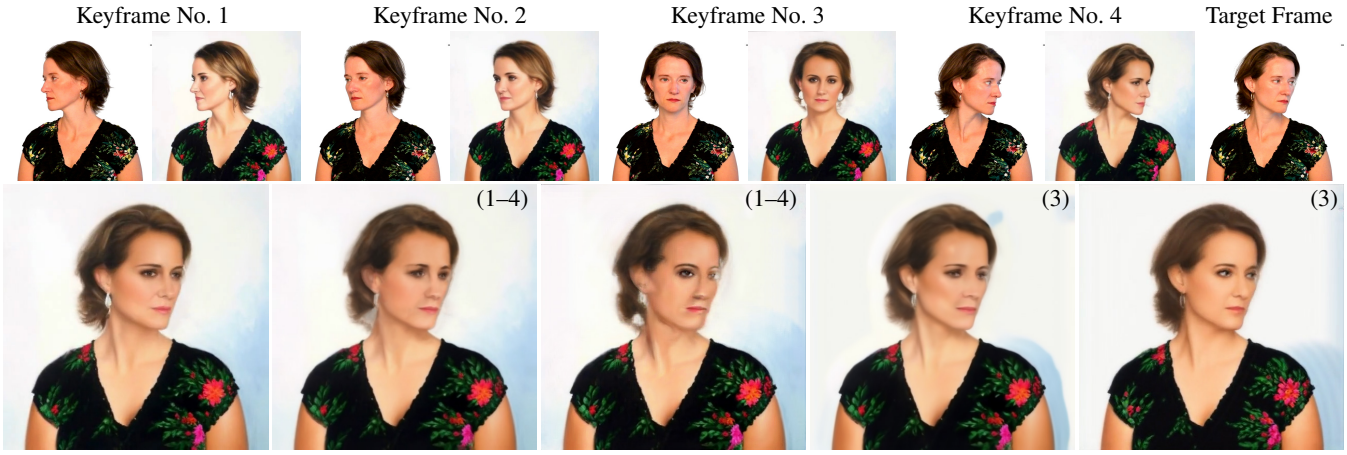


Figure S5: *Comparison with the state-of-the-art in text-driven video stylization (cont.): See Fig. S1 for detailed explanation.*

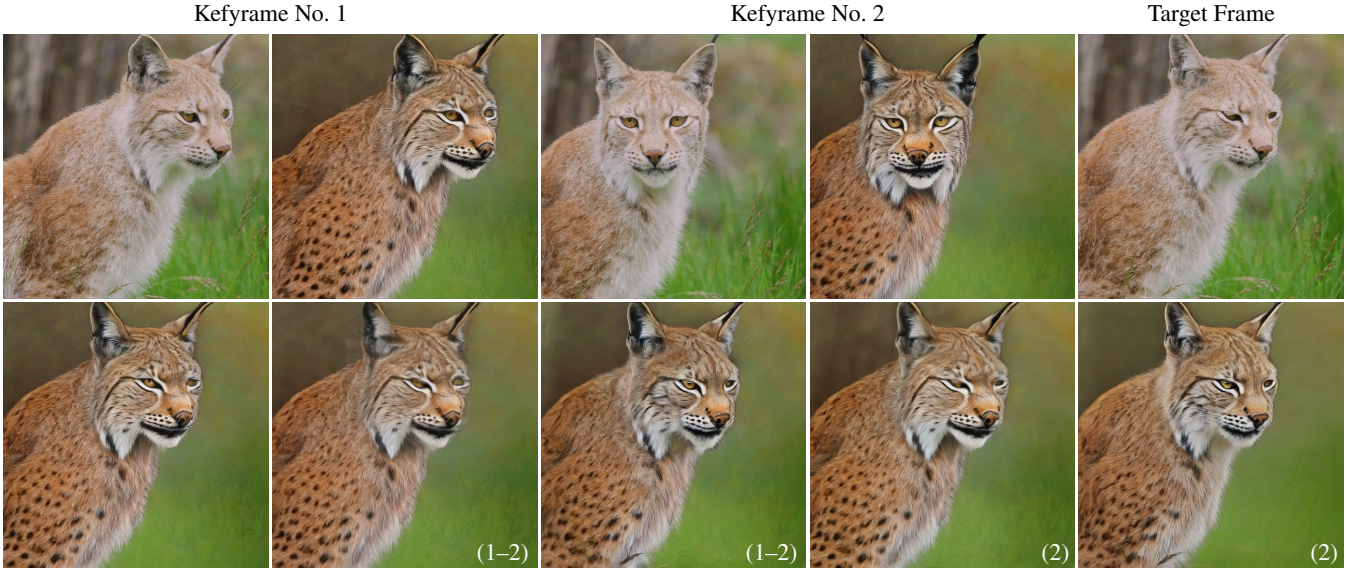
Figure S6: *Comparison with the state-of-the-art in text-driven video stylization (cont.): The target video sequence (a) has been stylized using text-driven approach of Yang et al. [YZLL24] (b). One frame from those stylized sequences was used as a keyframe (d). The style of this keyframe has been propagated to the rest of the target sequence* $\mathbf{y} \in \mathcal{Y}$ *using our approach (c). Note how our approach (c) better preserves the structural details seen in the target frame (a). Also, see our supplementary video to compare consistency across the entire sequence.*

© 2025 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.



(a) Yang et al. (b) Jamriška et al. (c) Texler et al. (d) Futschik et al. (e) Our Approach

Figure S7: *Comparison with the state-of-the-art in keyframe-based video stylization: The text-driven method of Yang et al. [YZLL23] has been used to generate a stylized sequence (a) from which four keyframes (No. 1–4) were selected to perform video stylization using methods of Jamriška et al. [JST*19] (b) and Texler et al. [TFK*20] (c), and one keyframe (No. 3) was selected for the method of Futschik et al. [FKL*21] (d) and for our approach (e). Note how our approach better preserves the structural details seen in the target frame. Our supplementary video further illustrates the stability of our results across the entire sequence.*



(a) Geyer et al. (b) Jamriška et al. (c) Texler et al. (d) Futschik et al. (e) Our Approach

Figure S8: *Comparison with the state-of-the-art in keyframe-based video stylization (cont.): Text-driven method of Geyer et al. [GBTBD24] has been used to generate a stylized sequence (a) and one keyframe (No. 2) was selected for the method of Futschik et al. [FKL*21] (d) and for our approach (e). See Fig. S7 for detailed explanation.*

© 2025 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.

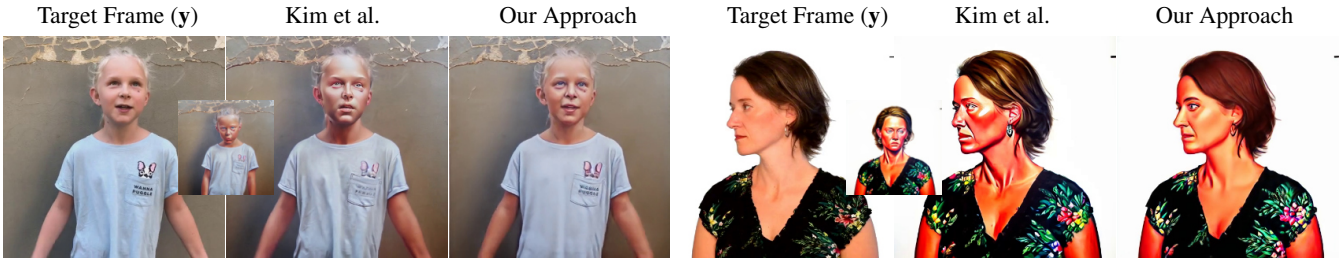


Figure S9: *Comparison with the state-of-the-art in text-driven video stylization (cont.): the method of Kim et al. [KLC*24] was used to generate a stylized sequence with the edit prompt: "hyperrealistic detailed oil painting of a girl/woman." From that sequence,* 1 *keyframe was selected (see small insets) to perform video stylization using our method. Note how our approach better preserves structural details seen in the target frame. In our supplementary video, you can see how our method outperforms the approach of Kim et al. with respect to the overall structural stability.*

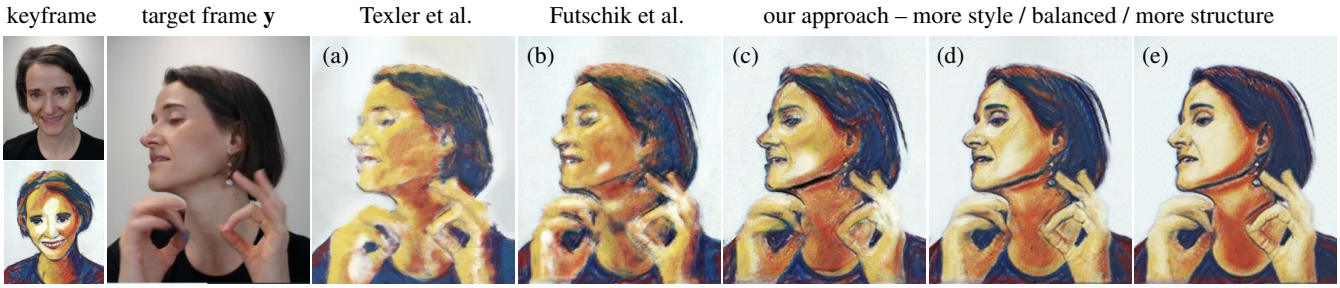


Figure S10: *Our approach applied in the real-time video call scenario originally proposed by Texler et al. [TFK*20] (a) and later improved by Futschik et al. [FKL*21] (b) in comparison with three different settings of our approach: (c) emphasizing style features from the stylized keyframe, (d) achieving a balance between style and structure, and (e) prioritizing structure from the target frame. All methods were trained with a single keyframe (left). As demonstrated in the supplementary video, our results show fewer flicker artifacts compared to other methods. The training time of our approach is comparable to the method of Futschik et al. [FKL*21].*

We find that point by assuming a standard logistic psychometric curve for a 2-AFC task:

$$P_{\text{correct}}(x) \;=\; 0.5 \;+\; 0.5\,\sigma(b_0 + b_1 x), \quad \sigma(z) = \frac{1}{1+e^{-z}}, \quad \text{(S1)}$$

where x is the *objective* distance between two methods ($+$ means our method is better), b_0 is a bias term, and b_1 is the slope. Given votes of participants (k_i, n_i) collected at distances x_i (k correct answers out of n trials), we estimate b_0, b_1 by maximising the binomial log-likelihood (equivalently minimising its negative):

$$\mathcal{L}(b_0, b_1) = -\sum_i \Big[k_i \ln P_{\text{correct}}(x_i) \;+\; (n_i - k_i) \ln\big(1 - P_{\text{correct}}(x_i)\big) \Big].$$

Because Eq. (S1) is monotonic, the 75 % threshold is reached when $P_{\text{correct}} = 0.75$, which occurs at

$$x_{75\,\%\,\text{JND}} \;=\; -\frac{b_0}{b_1}.$$

A **positive** $x_{75\,\%\,\text{JND}}$ tells us "how much the metric must favour *our* method before 75 % of viewers reliably prefer it."

Interpreting, for example, a style-related improvement of roughly $+0.11$ SSIM (or a *decrease* of 1.13 in LPIPS) is enough for 75 % of viewers to say that our stylized video looks better than the baseline. The smaller temporal-coherence thresholds show that viewers notice flicker with much finer metric changes than they need for overall style or structure.

Perceptual dimension	SSIM	LPIPS	ƎLIPS
Style	+0.1091	+1.1297	+0.6742
Structure	+0.1984	+0.1574	+0.1200
Temporal coherence	+0.1128	+0.0817	+0.0605

Table S1: *Estimated 75 % JND thresholds (metric units). Positive numbers mean the metric must improve by that amount in favour of our method before three-quarters of observers notice.*

We performed 75% JND threshold analysis with perceptual study presented in Sec. 4.1 and quantitative measurements presented in Sec. 4.2 of the main paper and we present the results in Table S1. Table 1 in the main paper demonstrates that our approach consistently outperforms previous methods *numerically*. However, the JND analysis clarifies *which* of those gaps matter perceptually – only the SSIM leap over Gen-3 Alpha crosses the just-noticeable line. Put differently, most of the smaller metric advantages we obtain over earlier research may still be invisible to the majority of

© 2025 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.

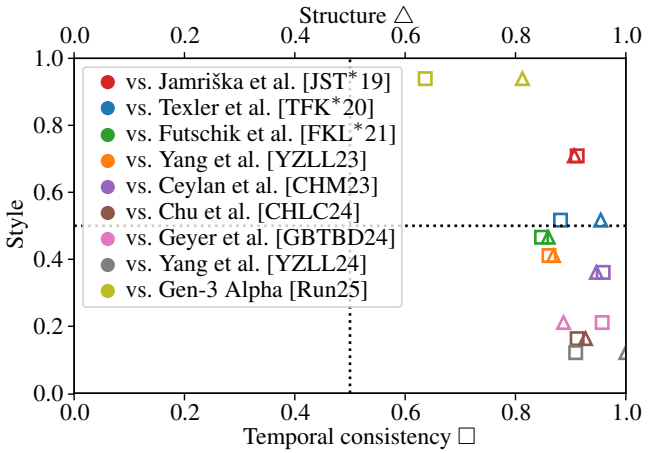


Figure S11: *Perceptual study. Each point represents the ratio of votes preferring the results of our method over those of other methods, based on responses from a total of* 55 *participants. Comparisons were made against three keyframe-based methods – Jamriška et al. [JST*19] (red), Texler et al. [TFK*20] (blue), and Futschik et al. [FKL*21] (green), five text-driven methods – Yang et al. [YZLL23, YZLL24] (orange, gray), Ceylan et al. [CHM23] (purple), Chu et al. [CHLC24] (brown), Geyer et al. [GBTBD24] (pink), and a large video model Gen-3 Alpha [Run25] (olive green). The bottom x-axis displays the ratio of answers favoring our method for preserving temporal consistency (hollow squares), the top x-axis shows the ratio favoring our method for structure preservation (hollow triangles), and the y-axis represents the style reproduction ratio. The graph illustrates that our approach offers improved performance over previous methods in reproducing input structures and maintaining temporal consistency. It is noteworthy that only three prior methods are preferred for style preservation in more than 75% of cases, an outcome that appears somewhat counterintuitive given our primary emphasis on structural fidelity.*

end users, highlighting the need for both automatic metrics *and* perceptual studies when evaluating high-fidelity stylisation.

References

[BHE23] Brooks T., Holynski A., Efros A. A.: InstructPix2Pix: Learning to follow image editing instructions. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (2023), pp. 18392–18402.

[CHLC24] Chu E., Huang T., Lin S.-Y., Chen J.-C.: MeDM: Mediating image diffusion models for video-to-video translation with temporal correspondence guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2024).

[CHM23] Ceylan D., Huang C.-H. P., Mitra N. J.: Pix2video: Video editing using image diffusion. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (2023), pp. 23206–23217.

[FKL*21] Futschik D., Kučera M., Lukáč M., Wang Z., Shechtman E., Sýkora D.: STALP: Style transfer with auxiliary limited pairing. *Computer Graphics Forum 40*, 2 (2021), 563–573.

[GBTBD24] Geyer M., Bar-Tal O., Bagon S., Dekel T.: TokenFlow: Consistent diffusion features for consistent video editing. In *Proceedings of International Conference on Learning Representations* (2024).

[JST*19] Jamriška O., Sochorová Š., Texler O., Lukáč M., Fišer J., Lu J., Shechtman E., Sýkora D.: Stylizing video by example. *ACM Transactions on Graphics 38*, 4 (2019), 107.

[KLC*24] Kim S., Lee K., Choi J. S., Jeong J., Sohn K., Shin J.: Collaborative score distillation for consistent visual editing. In *Advances in Neural Information Processing Systems* (2024).

[Run25] Runway: Gen-3 Alpha, 2025. URL: `https://runwayml.com/research/introducing-gen-3-alpha`.

[TFK*20] Texler O., Futschik D., Kučera M., Jamriška O., Sochorová Š., Chai M., Tulyakov S., Sýkora D.: Interactive video stylization using few-shot patch-based training. *ACM Transactions on Graphics 39*, 4 (2020), 73.

[YZLL23] Yang S., Zhou Y., Liu Z., Loy C. C.: Rerender A Video: Zero-shot text-guided video-to-video translation. In *SIGGRAPH Asia Conference Papers* (2023), p. 95.

[YZLL24] Yang S., Zhou Y., Liu Z., Loy C. C.: FRESCO: Spatial-Temporal Correspondence for Zero-Shot Video Translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2024), pp. 8703–8712.

© 2025 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.