CONSTANT: Towards High-Quality One-Shot Handwriting Generation with Patch Contrastive Enhancement and Style-Aware Quantization

Anh-Duy Le, Van-Linh Pham, Thanh-Nam Vo, Xuan Toan Mai, Tuan-Anh Tran

Viettel Artificial Intelligence and Data Services Center | Ho Chi Minh City University of Technology

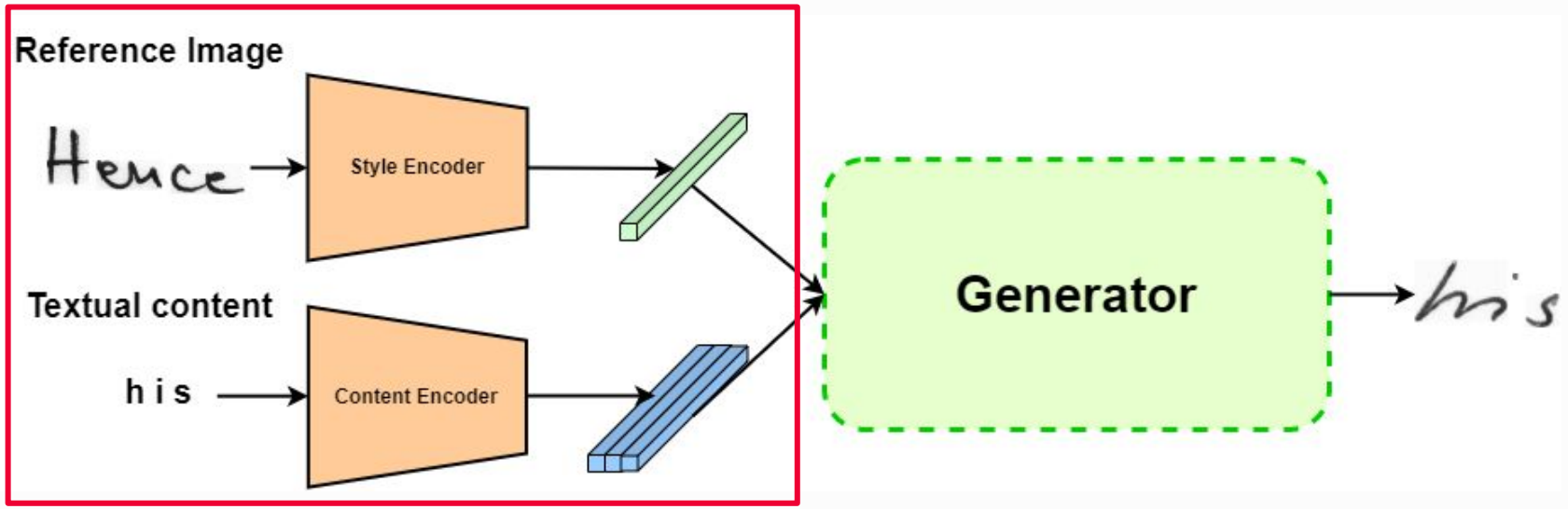
WACV | TUCSON, AZ | 2026

Background: Handwritten Text Generation

Handwritten Text Generation is the process of synthesizing new images of handwritten text that replicate a specific writing style.

Input:

- **Reference Style Image:** A single image provided by a writer that showcases their unique handwriting characteristics.
- **Textual Content:** The specific words or characters (text strings) that the user wants to be rendered in that handwriting style.



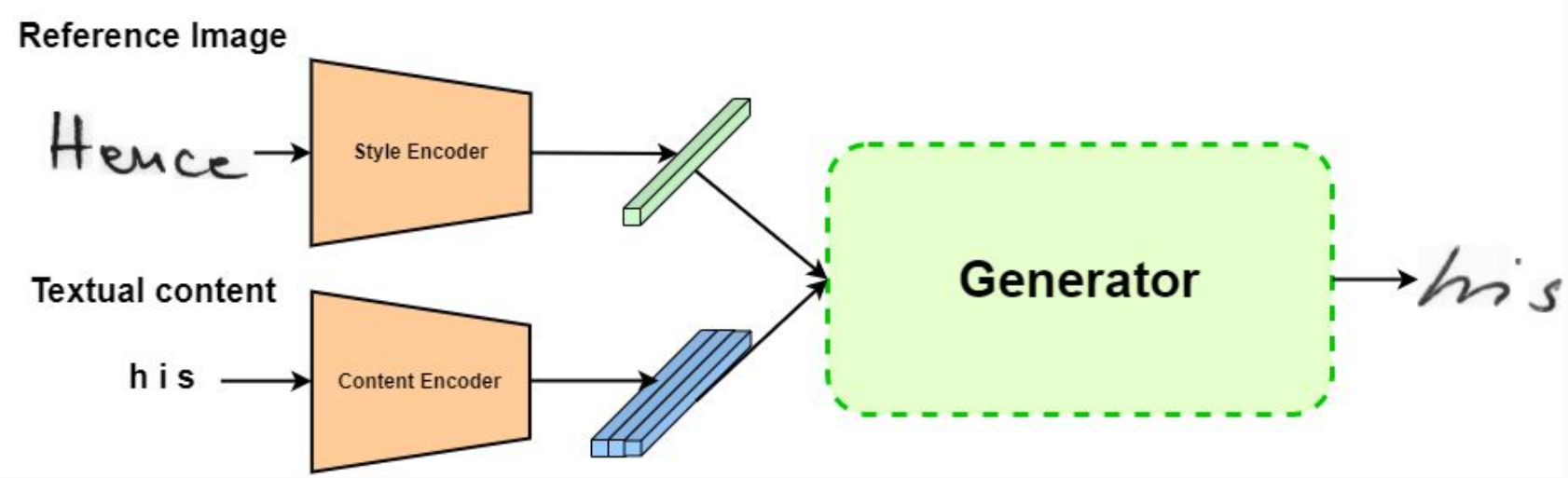
Background: Handwritten Text Generation

Handwritten Text Generation is the process of synthesizing new images of handwritten text that replicate a specific writing style.

Input:

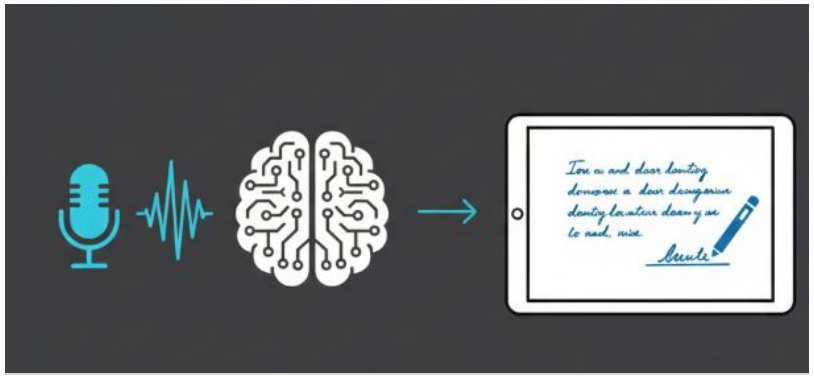
- **Reference Style Image:** A single image provided by a writer that showcases their unique handwriting characteristics.
- **Textual Content:** The specific words or characters (text strings) that the user wants to be rendered in that handwriting style.

Output: A realistic handwritten image that displays the input text while perfectly mimicking the style (e.g., slant, stroke width, ink density) of the reference image.



Background: Why HTG is important?

- **Assistive Technology**: Restores personal handwriting for individuals with motor impairments.
- **Data Augmentation**: Improves OCR accuracy by generating diverse, large-scale synthetic datasets.
- **Security & Forensics**: Strengthens handwriting-based authentication and forensic analysis testing.



Figures are generated by Google Nano Banana

Background: Why HTG is important?

- **Assistive Technology**: Restores personal handwriting for individuals with motor impairments.
- **Data Augmentation**: Improves OCR accuracy by generating diverse, large-scale synthetic datasets.
- **Security & Forensics**: Strengthens handwriting-based authentication and forensic analysis testing.

Figures are generated by Google Nano Banana

Background: Why HTG is important?

- **Assistive Technology**: Restores personal handwriting for individuals with motor impairments.
- **Data Augmentation**: Improves OCR accuracy by generating diverse, large-scale synthetic datasets.
- **Security & Forensics**: Strengthens handwriting-based authentication and forensic analysis testing.

Figures are generated by Google Nano Banana

Background: Why HTG is a challenging task ?

- **Style Complexity** Capturing the high variability of handwriting, such as unique slants, stroke widths, and subtle ink pressures, from a single sample.

- **Limited Generalization** Difficulty isolating core style features from background noise, leading to "incomplete" styles or failure on unseen, complex handwriting.

- **Content-Style Balance** The challenge of accurately replicating a writer's aesthetic while simultaneously ensuring the generated text is legible and correct.

Background: Why HTG is a challenging task ?

- **Style Complexity** Capturing the high variability of handwriting, such as unique slants, stroke widths, and subtle ink pressures, from a single sample.

- **Limited Generalization** Difficulty isolating core style features from background noise, leading to "incomplete" styles or failure on unseen, complex handwriting.

- **Content-Style Balance** The challenge of accurately replicating a writer's aesthetic while simultaneously ensuring the generated text is legible and correct.

Background: Why HTG is a challenging task ?

- **Style Complexity** Capturing the high variability of handwriting, such as unique slants, stroke widths, and subtle ink pressures, from a single sample.

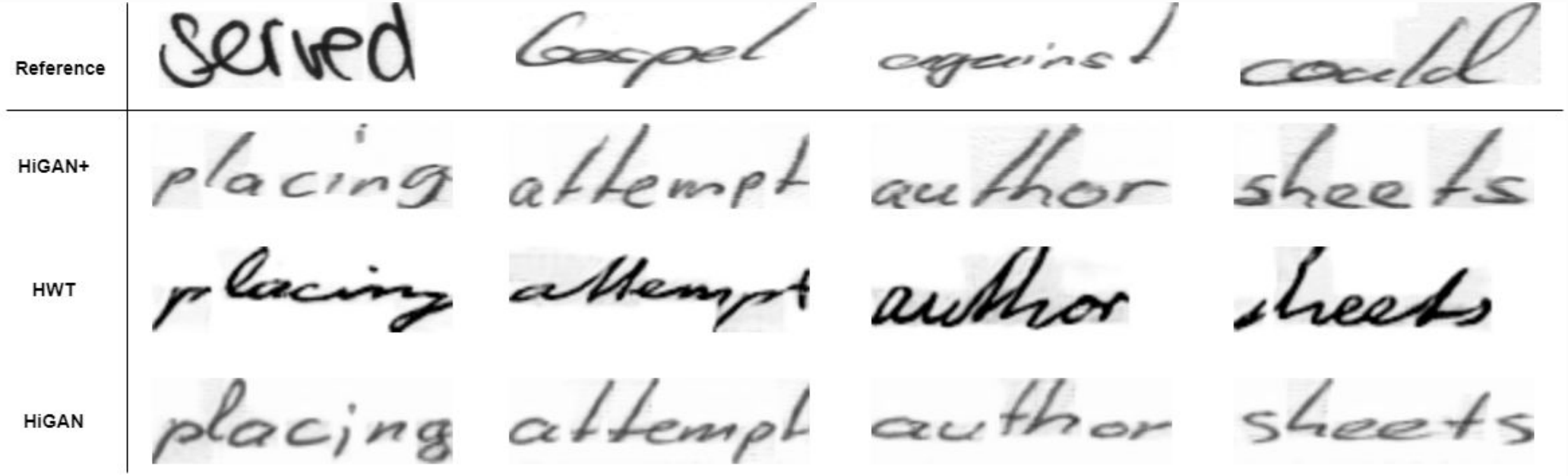
- **Limited Generalization** Difficulty isolating core style features from background noise, leading to "incomplete" styles or failure on unseen, complex handwriting.

- **Content-Style Balance** The challenge of accurately replicating a writer's aesthetic while simultaneously ensuring the generated text is legible and correct.

Prior Method Limitations

GAN-Based Approaches

- **Unstable training processes** hinder consistent output quality.
- **Struggles to generate realistic images** for complex writing styles.
- **Random noise inputs** prevent precise writer style control. (e.g. [1])

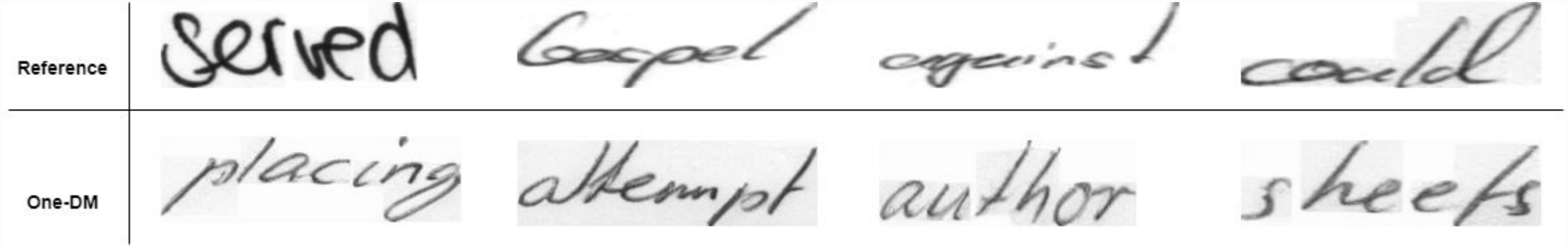


[1] Fogel, S., Averbuch-Elor, H., Cohen, S., Mazor, S., & Litman, R. (2020). Scrabblegan: Semi-supervised varying length handwritten text generation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 4324-4333).

Prior Method Limitations

Diffusion-Based Approaches

- **Inability to comprehensively model style** from one single image (e.g. One-DM[1]).
- **Predefined writer IDs** prevent generalization to unseen styles (e.g. WordStylist[2], CTIG-DM[3]).
- **Fails to produce correct textual content** and local details.
- **Standard denoising losses** lead to blurry, oversmoothed images (e.g. One-DM[1]).



[1] Dai, G., Zhang, Y., Ke, Q., Guo, Q., & Huang, S. (2024, September). One-dm: One-shot diffusion mimicker for handwritten text generation. In European Conference on Computer Vision (pp. 410-427). Cham: Springer Nature Switzerland.
[2] Nikolaidou, K., Retsinas, G., Christlein, V., Seuret, M., Sfikas, G., Smith, E. B., ... & Liwicki, M. (2023, August). Wordstylist: styled verbatim handwritten text generation with latent diffusion models. In International Conference on Document Analysis and Recognition (pp. 384-401). Cham: Springer Nature Switzerland.
[3] Zhu, Y., Li, Z., Wang, T., He, M., & Yao, C. (2023). Conditional text image generation with diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 14235-14245).

How can we effectively capture the complex stylistic characteristics from a single handwriting sample to generate diverse, realistic, and high-quality text image?

Key Ideas & Intuition

WACV 2026
TUCSON, AZ 3/6 - 3/10

Mirror Intuition

We aim to mirror human intuition by fundamentally categorizing specific writing traits, separating the core content from unique stylistic flourishes.

Discrete Tokens

Handwriting styles are modeled as discrete, noise-resistant visual tokens (style concepts) to create robust representations of unseen styles.

Contrastive Space

We create a discriminative latent space using contrastive learning and multi-scale patch objectives for sharper, high-fidelity local details.

Key Ideas & Intuition

WACV 2026 TUCSON, AZ 3/6 - 3/10

Mirror Intuition

We aim to mirror human intuition by fundamentally categorizing specific writing traits, separating the core content from unique stylistic flourishes.

Discrete Tokens

Handwriting styles are modeled as discrete, noise-resistant visual tokens (style concepts) to create robust representations of unseen styles.

Contrastive Space

We create a discriminative latent space using contrastive learning and multi-scale patch objectives for sharper, high-fidelity local details.

Key Ideas & Intuition

WACV 2026 TUCSON, AZ 3/6 - 3/10

Mirror Intuition

We aim to mirror human intuition by fundamentally categorizing specific writing traits, separating the core content from unique stylistic flourishes.

Discrete Tokens

Handwriting styles are modeled as discrete, noise-resistant visual tokens (style concepts) to create robust representations of unseen styles.

Contrastive Space

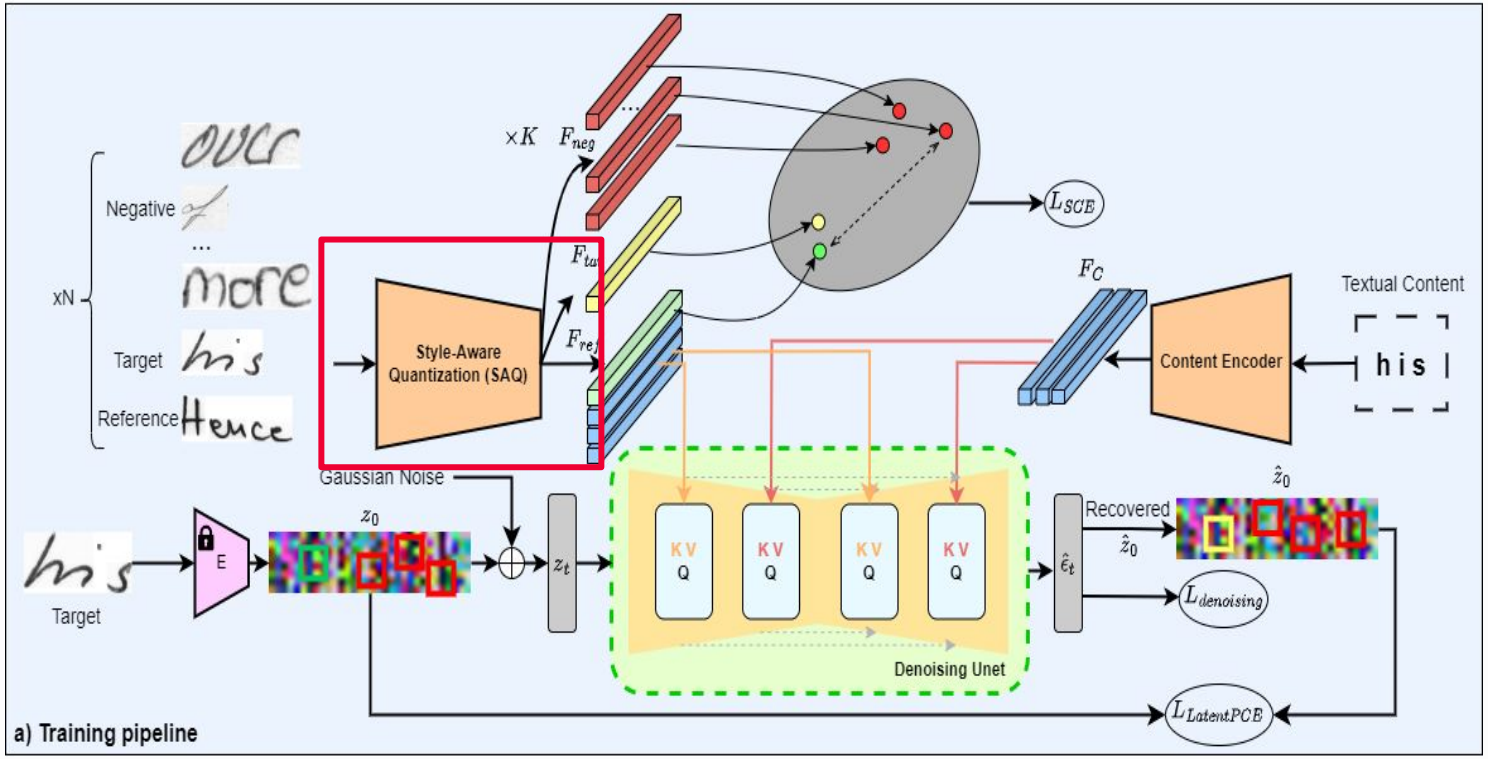
We create a discriminative latent space using contrastive learning and multi-scale patch objectives for sharper, high-fidelity local details.

CONSTANT Architecture

WACV 2026 TUCSON, AZ 3/6 - 3/10

Core components

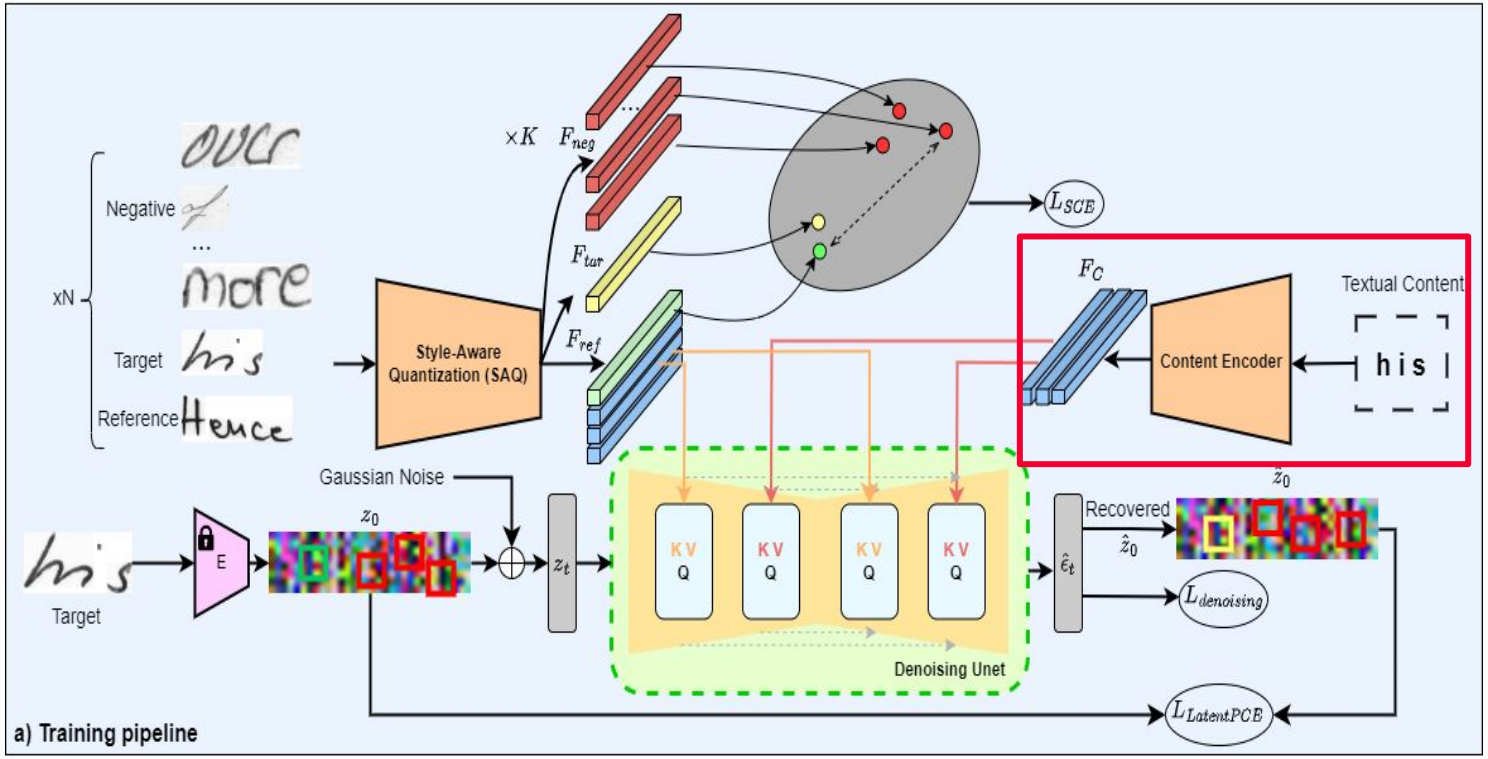
- **Style-Aware Quantization (SAQ)**: InceptionV3-based VQ module that discretizes style into noise-free visual concepts, with codebook optimized by L_SAQ.
- **Content Encoder**: 3-layer Transformer converting text into robust, character-level embeddings.
- **Latent Diffusion Model**: Latent-space UNet fusing style/content via cross-attention to denoise images.
- **Contrastive Objectives**: Global (L_SCE) and local (L_LatentPCE) losses that refine style and sharpen patch-level details.



CONSTANT Architecture

Core components

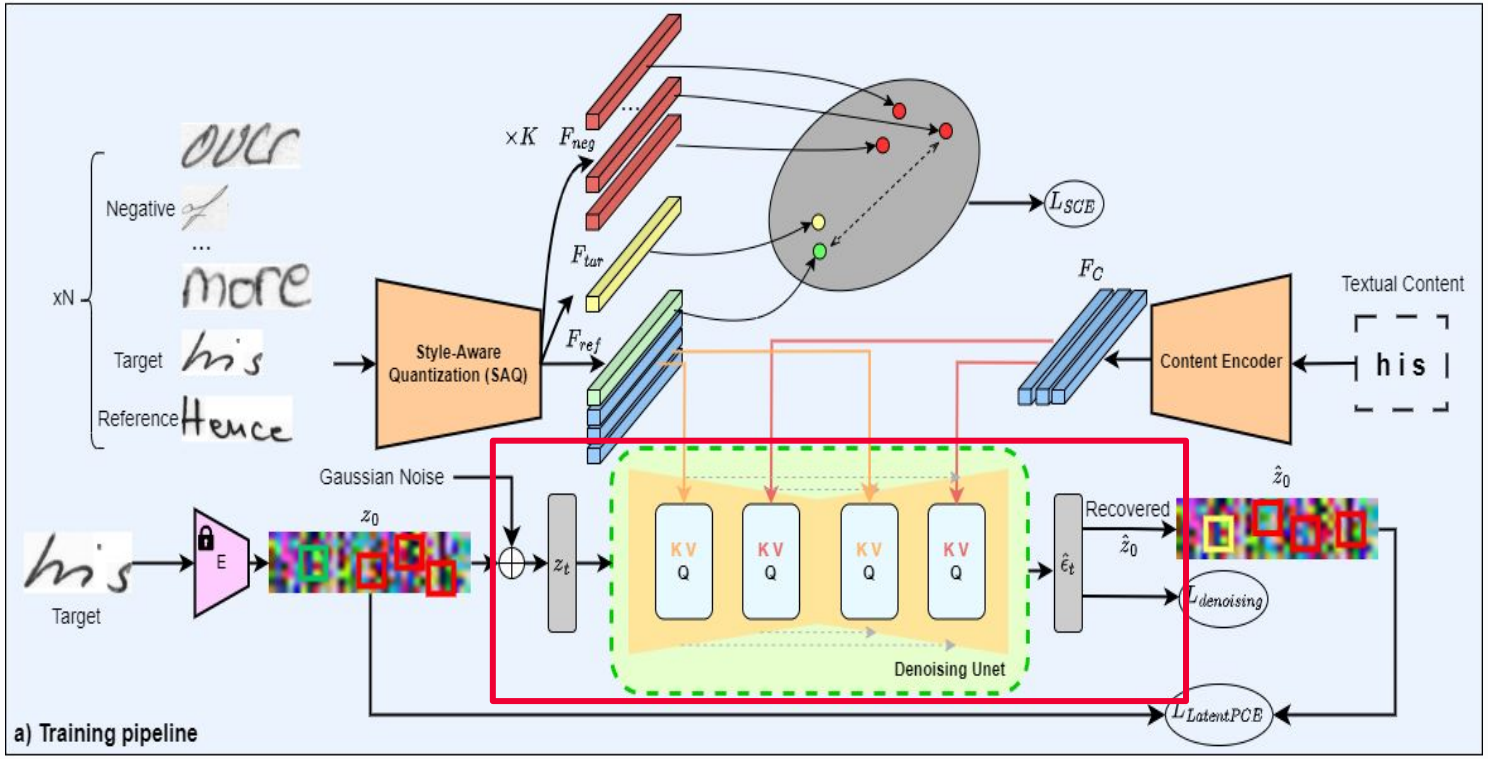
- **Style-Aware Quantization (SAQ)**: InceptionV3-based VQ module that discretizes style into noise-free visual concepts, with codebook optimized by L_SAQ. .
- **Content Encoder**: 3-layer Transformer converting text into robust, character-level embeddings.
- **Latent Diffusion Model**: Latent-space UNet fusing style/content via cross-attention to denoise images.
- **Contrastive Objectives**: Global (L_SCE) and local (L_LatentPCE) losses that refine style and sharpen patch-level details.



CONSTANT Architecture

Core components

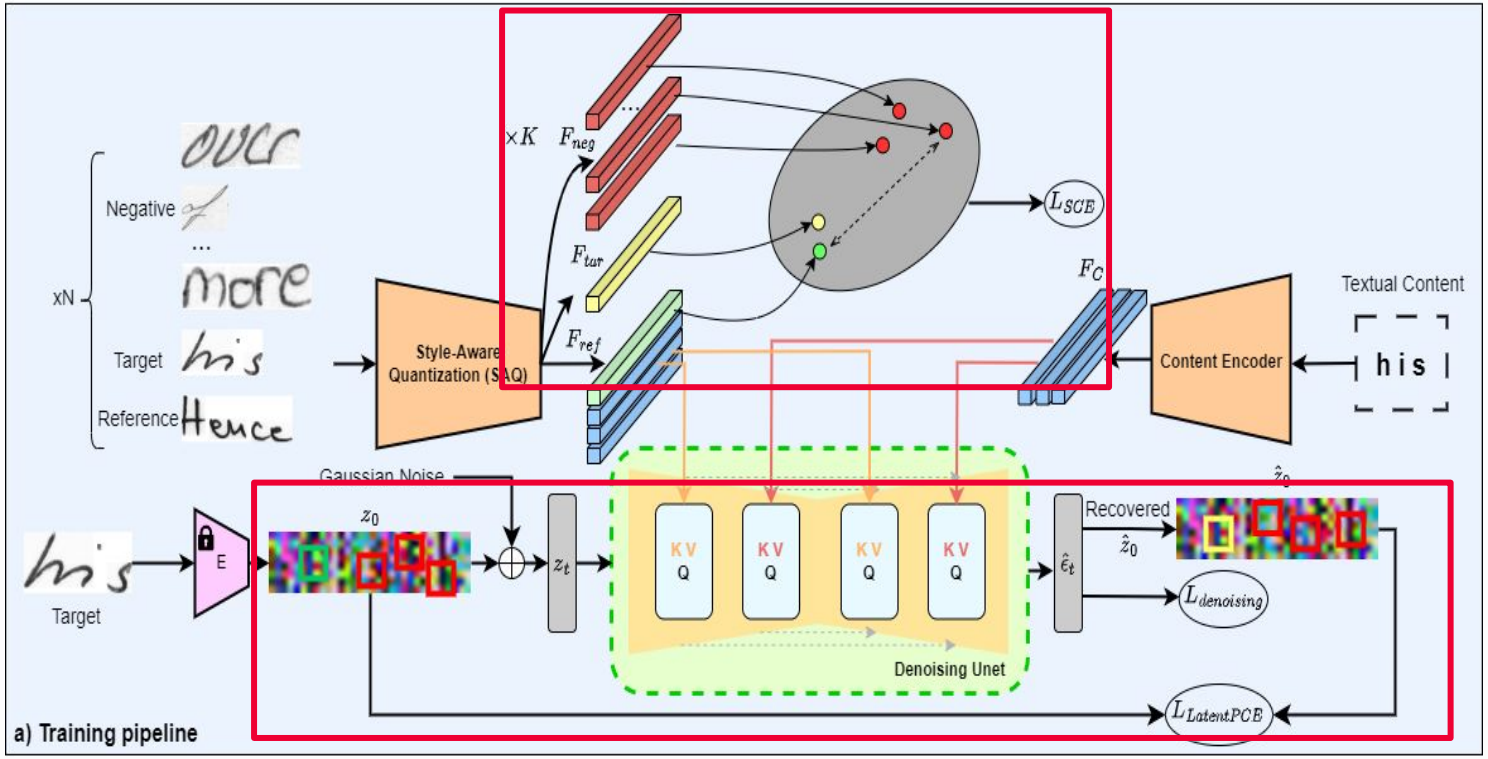
- **Style-Aware Quantization (SAQ)**: InceptionV3-based VQ module that discretizes style into noise-free visual concepts, with codebook optimized by L_SAQ. .
- **Content Encoder**: 3-layer Transformer converting text into robust, character-level embeddings.
- **Latent Diffusion Model**: Latent-space UNet fusing style/content via cross-attention to denoise images.
- **Contrastive Objectives**: Global (L_SCE) and local (L_LatentPCE) losses that refine style and sharpen patch-level details.



CONSTANT Architecture

Core components

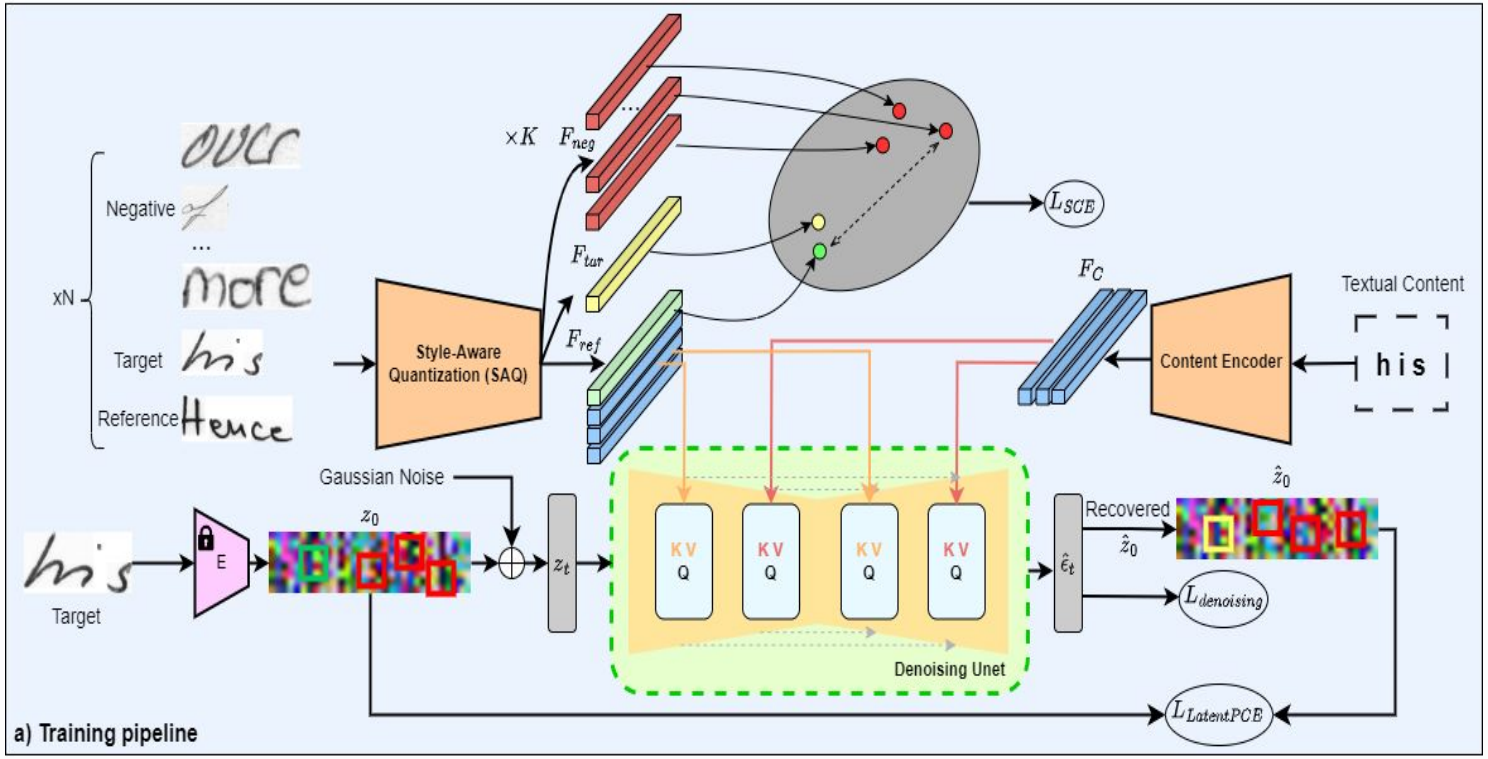
- **Style-Aware Quantization (SAQ)**: InceptionV3-based VQ module that discretizes style into noise-free visual concepts, with codebook optimized by L_SAQ. .
- **Content Encoder**: 3-layer Transformer converting text into robust, character-level embeddings.
- **Latent Diffusion Model**: Latent-space UNet fusing style/content via cross-attention to denoise images.
- **Contrastive Objectives**: Global (L_SCE) and local (L_LatentPCE) losses that refine style and sharpen patch-level details.



CONSTANT Architecture

Core components

- **Style-Aware Quantization (SAQ)**: InceptionV3-based VQ module that discretizes style into noise-free visual concepts.
- **Content Encoder**: 3-layer Transformer converting text into robust, character-level embeddings.
- **Latent Diffusion Model**: Latent-space UNet fusing style/content via cross-attention to denoise images.
- **Contrastive Objectives**: Global (L_SCE) and local (L_LatentPCE) losses that refine style and sharpen patch-level details.



Overall Objective $L = L_{denoising} + \alpha \times (L_{LatentPCE} + L_{SCE} + L_{SAQ})$

Methodology: Style-Aware Quantization (SAQ)

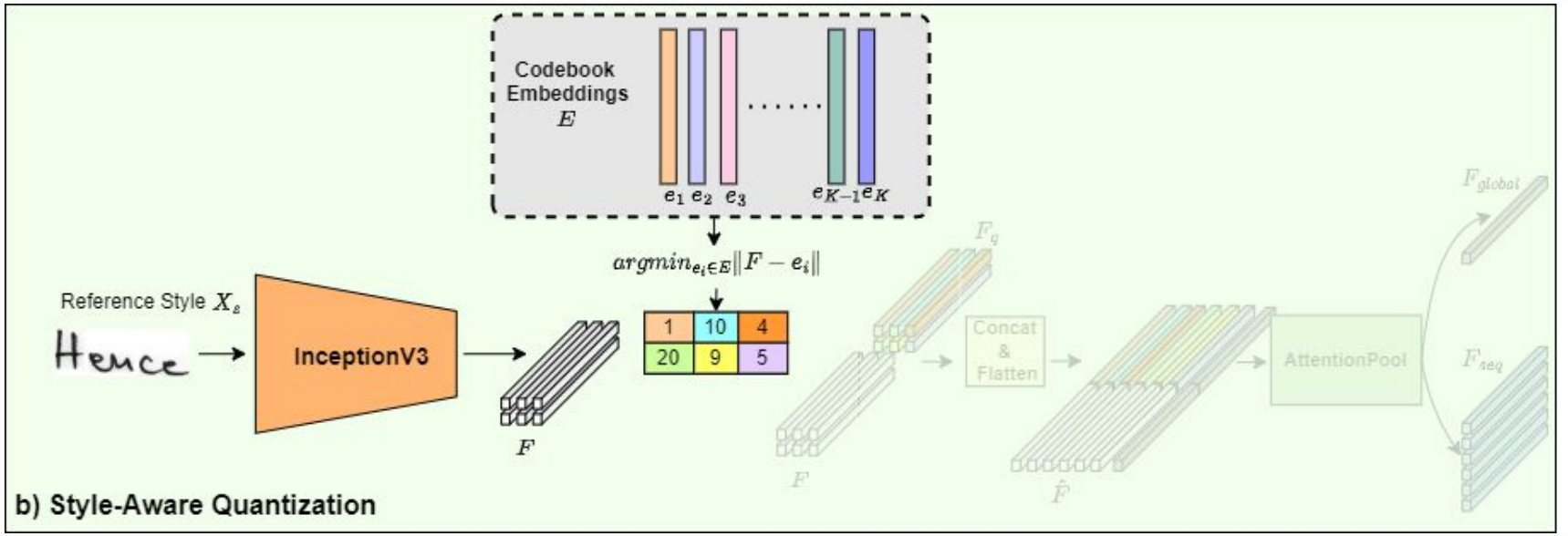
WACV 2026 TUCSON, AZ 3/6 - 3/10

Discretization

The SAQ maps continuous visual features to a codebook of discrete "style concepts". This Vector Quantization (VQ) approach is highly effective at filtering out irrelevant noise while retaining the core stylistic essence of the unseen writer.

Hybrid Fusion

It concatenates the continuous and quantized features, processing them through an Attention Pool module. This outputs a global representation F_{global} for writer discrimination and a sequence of fused features F_{seq} as vital context for the diffusion model.



Methodology: Style-Aware Quantization (SAQ)

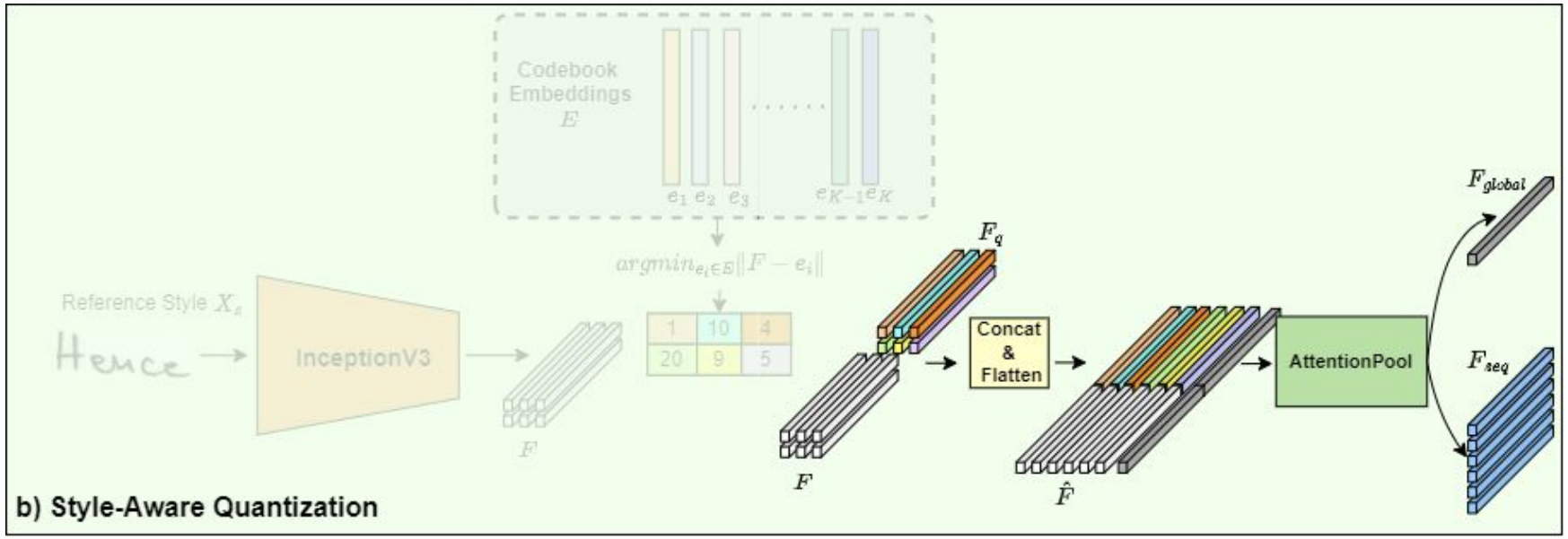
WACV 2026 TUCSON, AZ 3/6 - 3/10

Discretization

The SAQ maps continuous visual features to a codebook of discrete "style concepts". This Vector Quantization (VQ) approach is highly effective at filtering out irrelevant noise while retaining the core stylistic essence of the unseen writer.

Hybrid Fusion

It concatenates the continuous and quantized features, processing them through an Attention Pool module. This outputs a global representation F_{global} for writer discrimination and a sequence of fused features F_{seq} as vital context for the diffusion model.



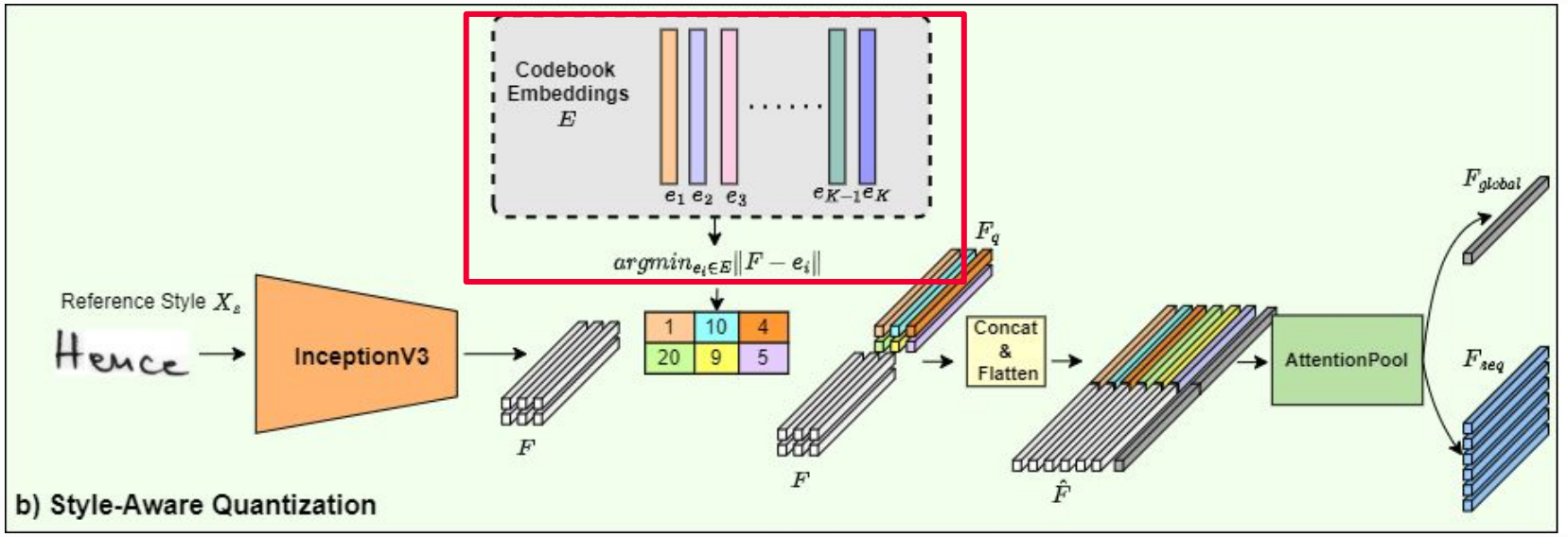
Methodology: Style-Aware Quantization (SAQ)

Similar to previous work [1]:

VQ Loss: Aligns codebook embeddings with extracted handwriting features.

Commitment Loss: Ensures encoder features stay near learned style tokens.

Optimization: L_SAQ = L_vq + 0.25 x L_cmt stabilizes the discrete space.



[1] Van Den Oord, A., & Vinyals, O. (2017). Neural discrete representation learning. Advances in neural information processing systems, 30.

Methodology: Style Contrastive Enhancement

WACV 2026 TUCSON, AZ 3/6 - 3/10

L_{SCE}

Refining Latent Space

This objective utilizes the global style feature from the Attention Pool module. It is designed to pull reference and target styles of the **same** writer closer together, while pushing features of **different** writers further apart.

Discriminative Embedding

By swapping roles of the reference and target as different views in a bidirectional format, the framework ensures a highly discriminative embedding space that prevents the style encoder from simply memorizing the input image.

Methodology: Style Contrastive Enhancement

L_{SCE}

Refining Latent Space

This objective utilizes the global style feature from the Attention Pool module. It is designed to pull reference and target styles of the **same** writer closer together, while pushing features of **different** writers further apart.

Discriminative Embedding

By swapping roles of the reference and target as different views in a bidirectional format, the framework ensures a highly discriminative embedding space that prevents the style encoder from simply memorizing the input image.

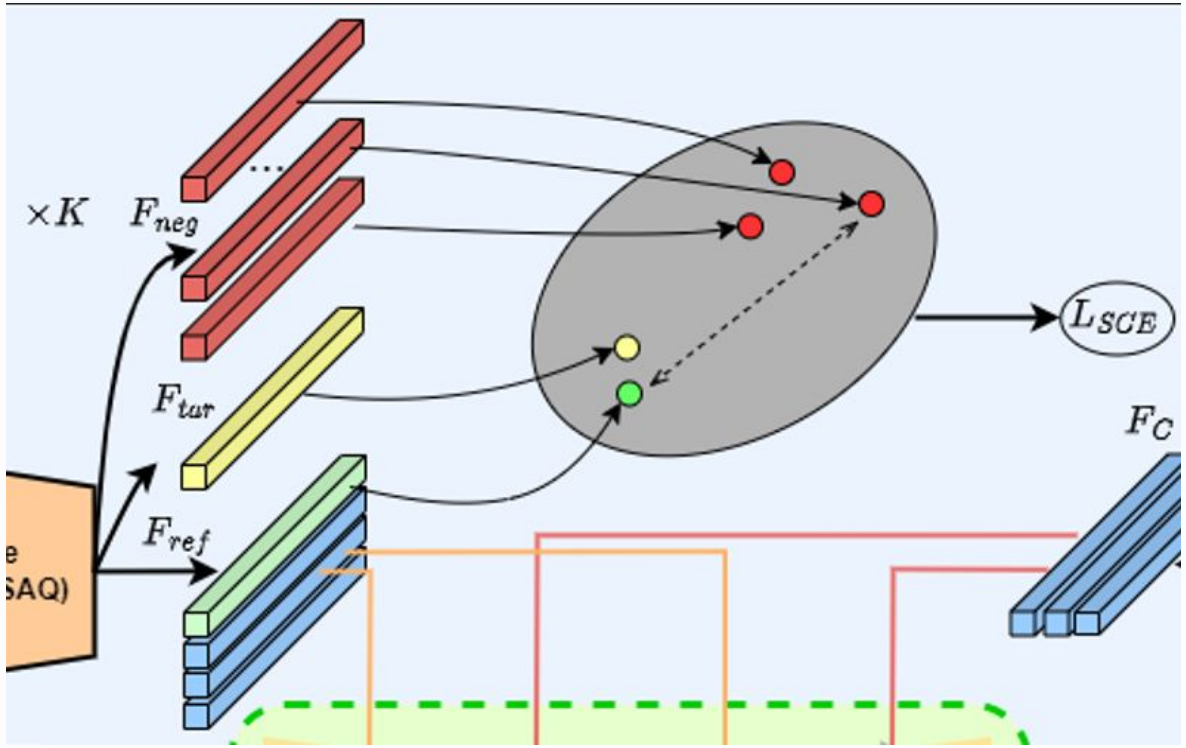
Methodology: Style Contrastive Enhancement

$$\ell_{SCE}\left(F_{tar}, F_{ref}, F_d\right) = \frac{-1}{N}\sum_{t\in B}\log\frac{\exp\left(\mathrm{sim}\left(F_{tar}, F_{ref}\right)/\tau\right)}{\exp\left(\mathrm{sim}\left(F_{tar}, F_{ref}\right)/\tau\right) + \sum_{d\in D(t)}\exp\left(\mathrm{sim}\left(F_{tar}, F_d\right)/\tau\right)}$$

Model optimizes the loss from both the target and reference perspectives using a bidirectional format

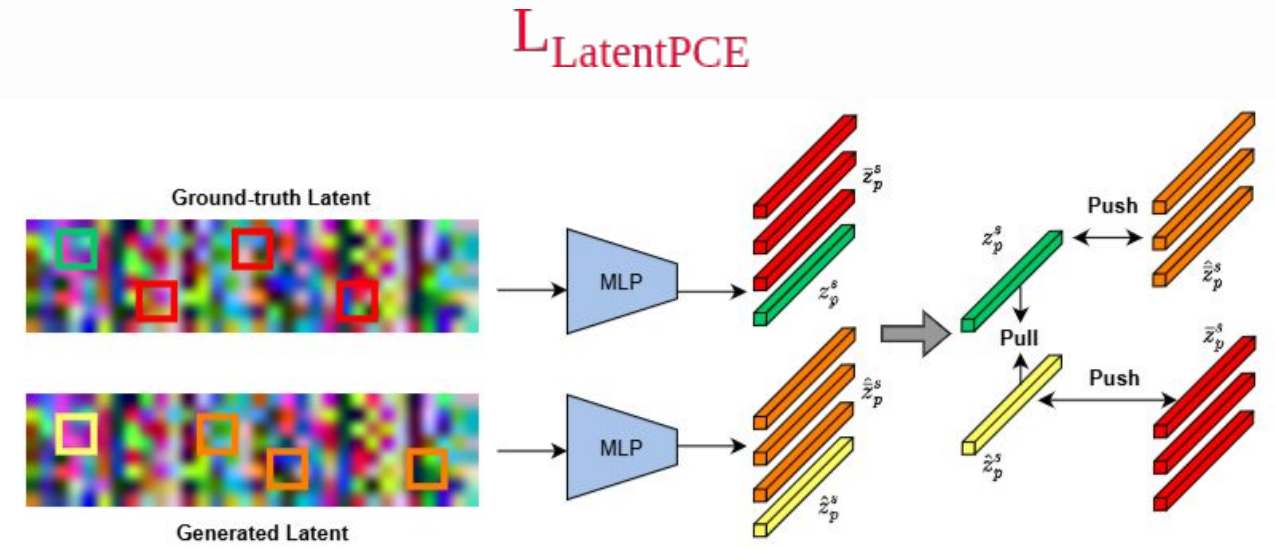
$$L_{SCE} = \frac{1}{2}\ell_{SCE}\left(F_{tar}, F_{ref}, \mathrm{sg}\left(F_d\right)\right) + \frac{1}{2}\ell_{SCE}\left(F_{ref}, F_{tar}, \mathrm{sg}\left(F_d\right)\right)$$

Stop-Gradient (sg): Prevents unstable gradients from passing through negative samples follow [1].



[1] Chen, X., & He, K. (2021). Exploring simple siamese representation learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 15750-15758).

Methodology: Latent Patch Contrastive Enhancement



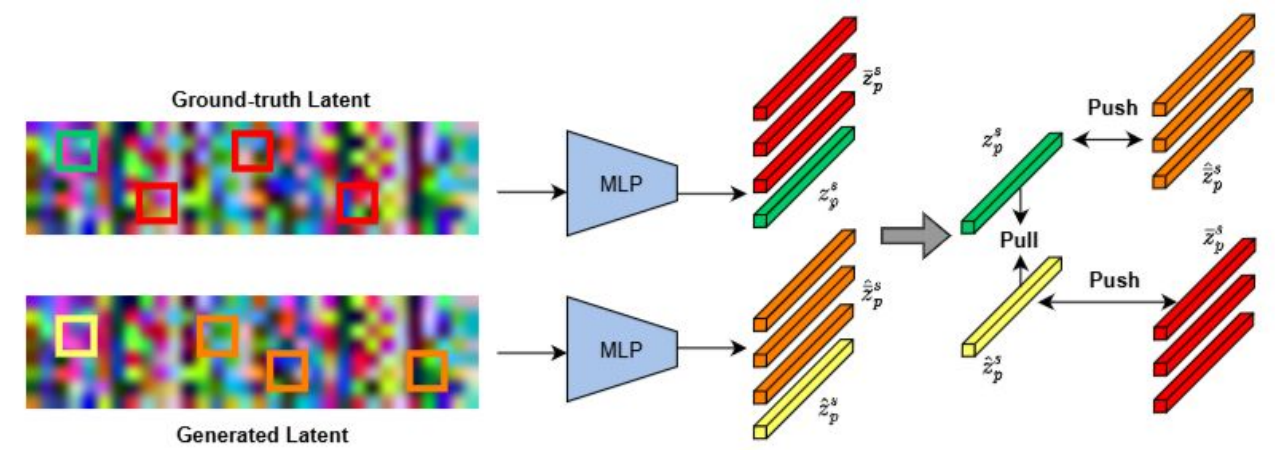
Sharpening Stroke Details

Latent Space Operation: Directly optimizes locality information within the diffusion model's latent space, rather than relying on pixel-level operations.

Spatial Alignment & Multi-Scale: Aligns patches from the same spatial locations of generated and ground-truth features. Captures details across different resolutions by maximizing mutual information.

Methodology: Latent Patch Contrastive Enhancement

$L_{LatentPCE}$



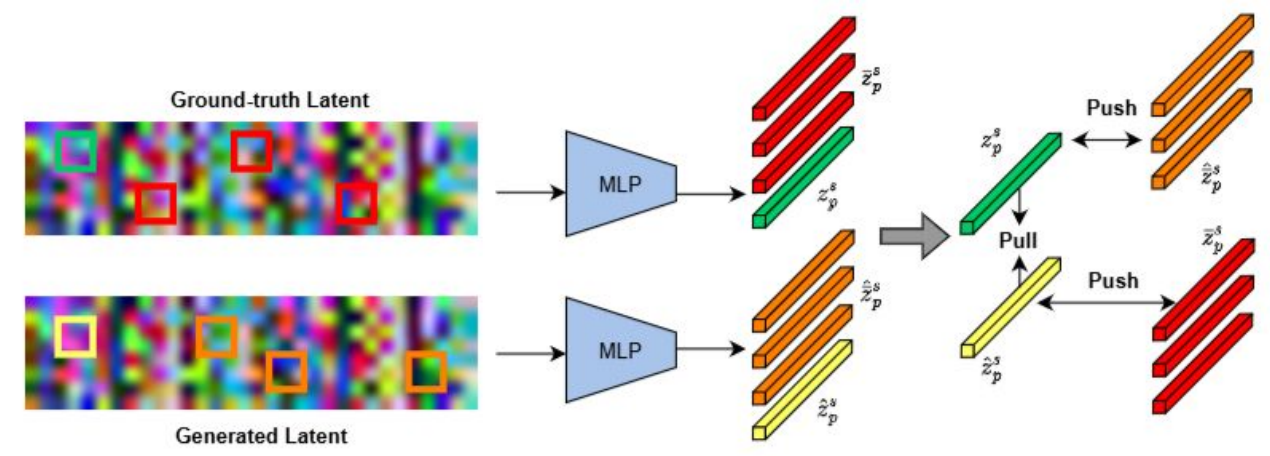
Sharpening Stroke Details

Latent Space Operation: Directly optimizes locality information within the diffusion model's latent space, rather than relying on pixel-level operations.

Spatial Alignment & Multi-Scale: Aligns patches from the same spatial locations of generated and ground-truth features. Captures details across different resolutions by maximizing mutual information.

Methodology: Latent Patch Contrastive Enhancement

- **Multi-scale Framework**: Operates across **three distinct resolutions** to capture varied levels of detail.
- **Patch Sampling**: Extracts **256 patches** per scale with sizes **2x2**, **4x4**, and **8x8**.
- **Feature Projection**: Patches are flattened and mapped to a **256-dimensional embedding** via a shallow MLP.



$$\ell_{LatentPCE}(z_p^s, \hat{z}_p^s, \hat{\bar{z}}_p^s) = -\frac{1}{H}\sum_{p=1}^{H} \log \frac{\exp\left(sim\left(z_p^s, \hat{z}_p^s\right)/\tau\right)}{\exp\left(sim\left(z_p^s, \hat{z}_p^s\right)/\tau\right) + \sum_{n=1}^{H_s^{-1}} \exp\left(sim\left(z_p^s, \hat{\bar{z}}_p^s\right)/\tau\right)}$$

Similar to LSCE , we can write L_LatentPCE in a bidirectional format

$$L_{LatentPCE} = \frac{1}{2S}\sum_{s=1}^{S} \ell_{LatentPCE}(z_p^s, \hat{z}_p^s, \mathrm{sg}(\hat{\bar{z}}_p^s)) + \frac{1}{2S}\ell_{LatentPCE}(\hat{z}_p^s, z_p^s, \mathrm{sg}(\bar{z}_p^s))$$

Experiment: Setting

Datasets:

IAM (500 writers), IMGUR5K (135,375 words and 5,305 diverse handwriting styles),

IIIT-English-Word (700,000 samples from 1,215 writers).

Evaluation metrics:

Visual Quality: **FID** (Fréchet Inception Distance) to measure realism.

Style Imitation: **HWD** (Handwriting Distance [1]) for geometric features and **Writer Classification Accuracy** (Acc_Wid).

Readability: **WER** (Word Error Rate) using a pretrained text recognition model [2].

[1] Pippi, V., Quattrini, F., Cascianelli, S., & Cucchiara, R. (2023). HWD: A novel evaluation score for styled handwritten text generation. arXiv preprint arXiv:2310.20316.

[2] Kass, D., & Vats, E. (2022, May). AttentionHTR: Handwritten text recognition based on attention encoder-decoder networks. In International Workshop on Document Analysis Systems (pp. 507-522). Cham: Springer International Publishing.

State-of-the-art Performance

WACV 2026 TUCSON, AZ 3/6 - 3/10

Method	Reference	HWD ↓	FID ↓	WER ↓	Acc_{Wid} ↑
HWT [2]	Few-shot	1.23	19.82	0.62	9.08
VATr [36]		1.13	16.30	0.51	49.81
DiffusionPen [33]		1.04	18.94	0.23	38.86
HiGAN+ [12]	One-shot	0.89	13.90	0.56	55.20
HiGAN [11]		1.55	27.13	0.55	29.79
One-DM [4]		1.05	15.97	0.36	4.5
Ours		**0.74**	**10.20**	**0.22**	**69.43**

Method	IV-S		OOV-S		IV-U		OOV-U	
	HWD ↓	FID ↓	HWD ↓	FID ↓	HWD ↓	FID ↓	HWD ↓	FID ↓
HWT [2]	2.30	135.51	2.30	146.16	2.35	138.39	2.36	148.75
VATr [36]	2.50	132.87	2.51	140.56	2.58	137.34	2.60	144.02
DiffusionPen [33]	1.14	91.20	1.16	97.65	1.52	112.87	1.65	122.52
HiGAN [11]	2.29	118.70	2.31	128.49	2.36	119.56	2.37	128.60
HiGAN+ [12]	1.76	117.76	1.79	122.56	1.78	117.63	1.81	124.38
One-DM [4]	1.95	104.04	1.99	107.81	1.94	117.74	1.99	121.94
Ours (CONSTANT)	**0.96**	**89.88**	**0.94**	**96.13**	**1.61**	**112.03**	**1.63**	**118.10**

Unrivaled SOTA: Achieves best-in-class scores across all metrics

State-of-the-art Performance

WACV 2026 TUCSON, AZ 3/6 - 3/10

Method	Reference	HWD ↓	FID ↓	WER ↓	Acc_{Wid} ↑
HWT [2]	Few-shot	1.23	19.82	0.62	9.08
VATr [36]		1.13	16.30	0.51	49.81
DiffusionPen [33]		1.04	18.94	0.23	38.86
HiGAN+ [12]	One-shot	0.89	13.90	0.56	55.20
HiGAN [11]		1.55	27.13	0.55	29.79
One-DM [4]		1.05	15.97	0.36	4.5
Ours		**0.74**	**10.20**	**0.22**	**69.43**

Method	IV-S		OOV-S		IV-U		OOV-U	
	HWD ↓	FID ↓	HWD ↓	FID ↓	HWD ↓	FID ↓	HWD ↓	FID ↓
HWT [2]	2.30	135.51	2.30	146.16	2.35	138.39	2.36	148.75
VATr [36]	2.50	132.87	2.51	140.56	2.58	137.34	2.60	144.02
DiffusionPen [33]	1.14	91.20	1.16	97.65	1.52	112.87	1.65	122.52
HiGAN [11]	2.29	118.70	2.31	128.49	2.36	119.56	2.37	128.60
HiGAN+ [12]	1.76	117.76	1.79	122.56	1.78	117.63	1.81	124.38
One-DM [4]	1.95	104.04	1.99	107.81	1.94	117.74	1.99	121.94
Ours (CONSTANT)	**0.96**	**89.88**	**0.94**	**96.13**	**1.61**	**112.03**	**1.63**	**118.10**

Unrivaled SOTA: Achieves best-in-class scores across all metrics

Robust Generalization: Dominates in all categories (**IV-S, OOV-S, IV-U, OOV-U**), maintaining high realism even for unseen styles and words.

Ablation Studies: Component Validation

Base	SAQ	L_{SCE}	L_{PCE}	Style images		FID ↓	HWD ↓
							
✓						16.73	0.87
✓	✓					12.47	0.85
✓	✓	✓				12.55	0.84
✓	✓	✓	✓			**10.20**	**0.74**

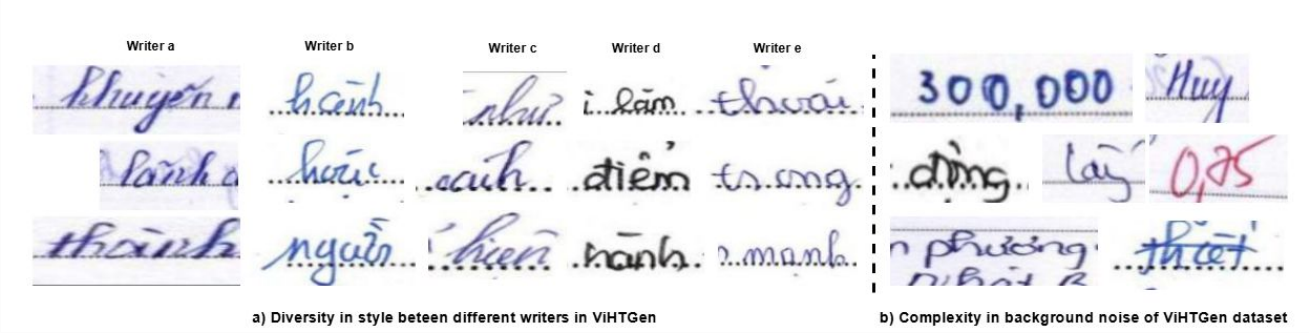
Each module contributes significantly. SAQ drastically improves general visual quality (FID), L_SCE improves style fidelity (HWD), and L_LatentPCE provides the final polish on local details.

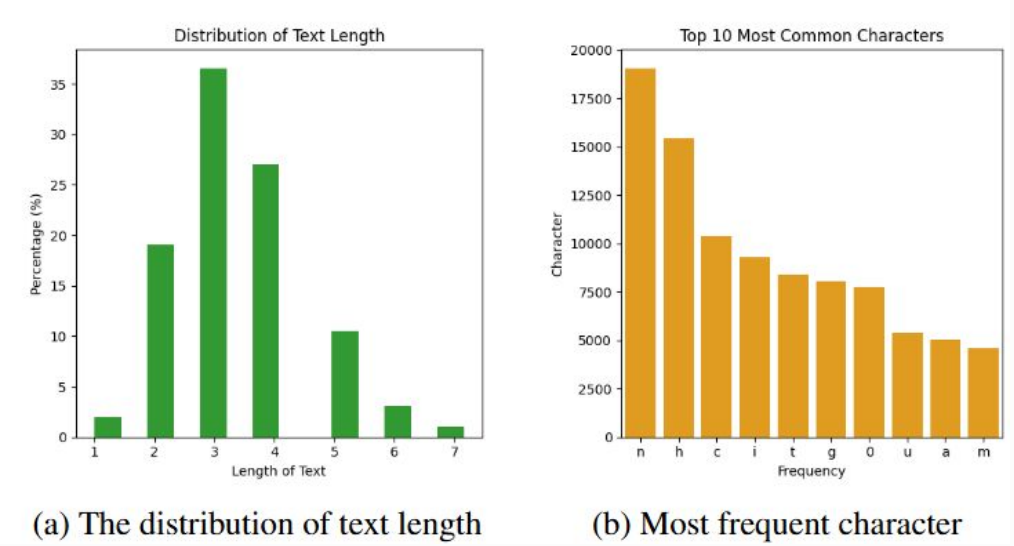
The Proposed ViHTGen Dataset

A Novel Benchmark: Features over 50,000+ Vietnamese handwritten words sourced from more than 300 university exam scripts.

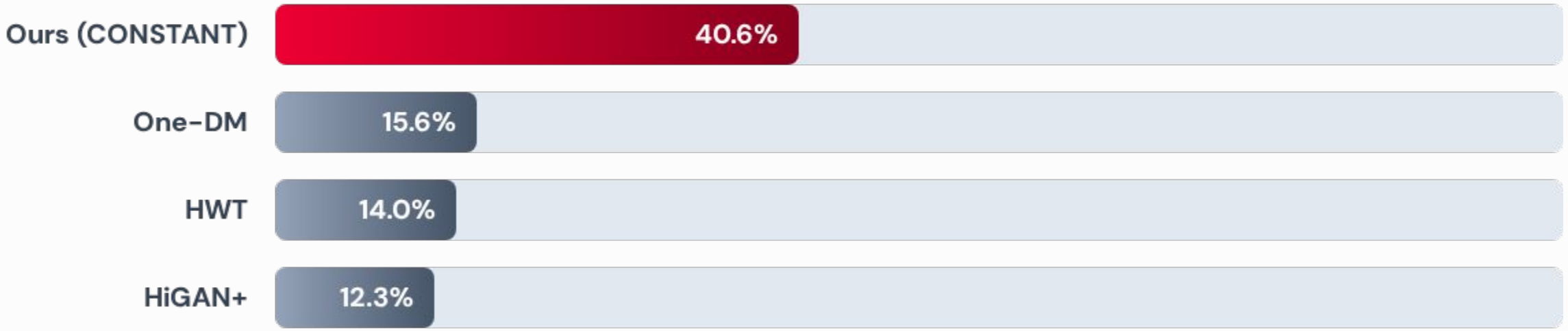
High Complexity: Unlike IAM, it features highly variable stroke widths, erratic slants, and complex, noisy backgrounds, making it a rigorous test for style generalization.

Domain Diversity: Includes text, numbers, and diverse ink colors. The ground truth was transcribed through a rigorous semi-automatic pipeline using VinTernVL-1B.





User Preference Study & Readability



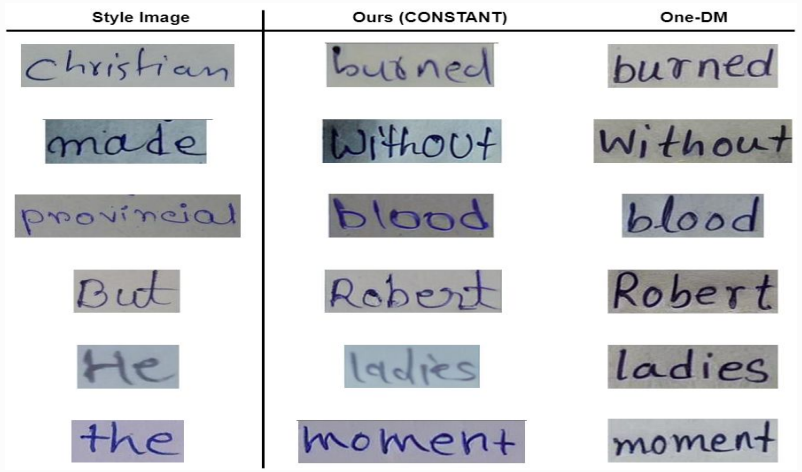
In a rigorous study with 840 responses, users selected CONSTANT over 40% of the time as the most visually realistic. Furthermore, training an OCR model on CONSTANT's generated data improved real-world recognition accuracy by 1.2%.

Visualizations & Qualitative Results

WACV 2026 TUCSON, AZ 3/6 - 3/10

Complex Cursive

Capturing fluid stroke dynamics and complex ligatures from one-shot samples.



Multilingual Support

Outperforming SOTA by 10% on Chinese and Vietnamese structural generation.

IMGUR5K Diversity

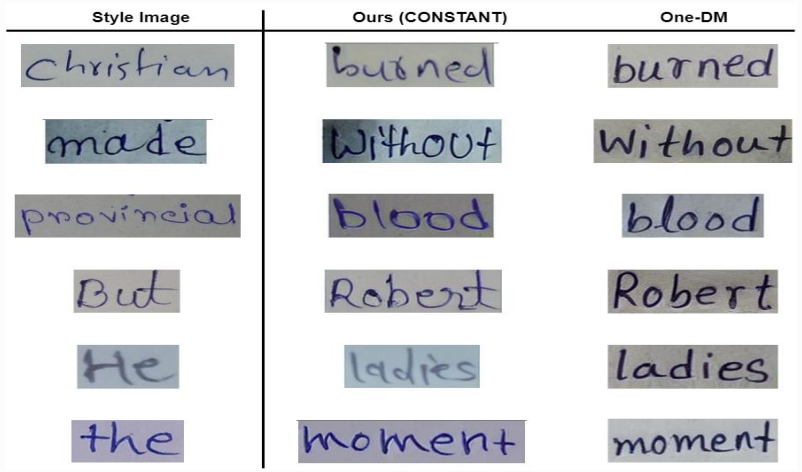
Achieving superior 0.99 HWD on highly diverse, real-world data distributions.

Visualizations & Qualitative Results

WACV 2026
TUCSON, AZ
3/6 - 3/10

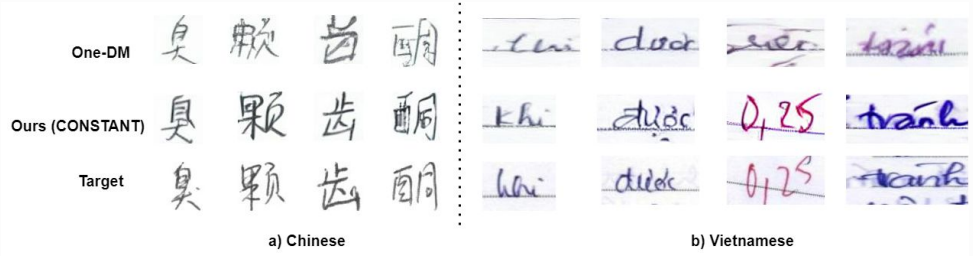
Complex Cursive

Capturing fluid stroke dynamics and complex ligatures from one-shot samples.



Multilingual Support

Outperforming SOTA by 10% on Chinese and Vietnamese structural generation.



IMGUR5K Diversity

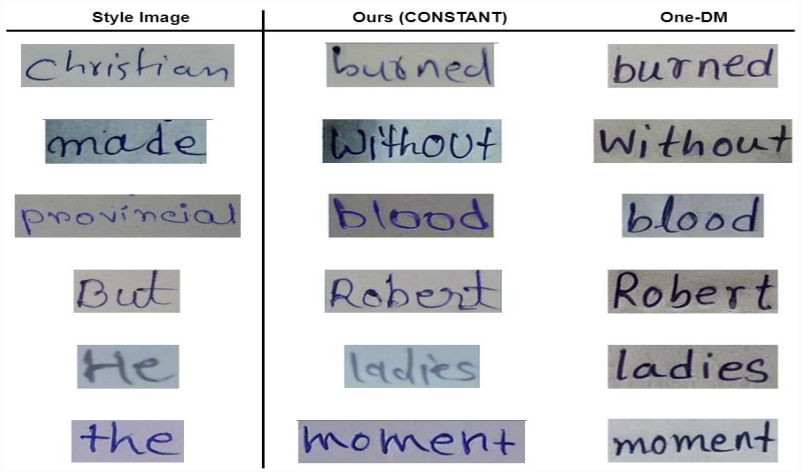
Achieving superior 0.99 HWD on highly diverse, real-world data distributions.

Visualizations & Qualitative Results

WACV 2026 TUCSON, AZ 3/6 - 3/10

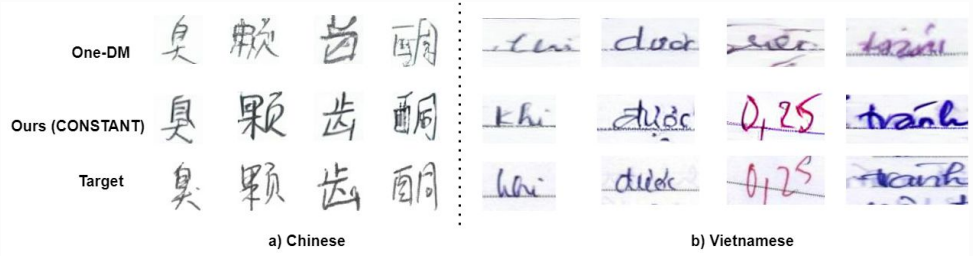
Complex Cursive

Capturing fluid stroke dynamics and complex ligatures from one-shot samples.



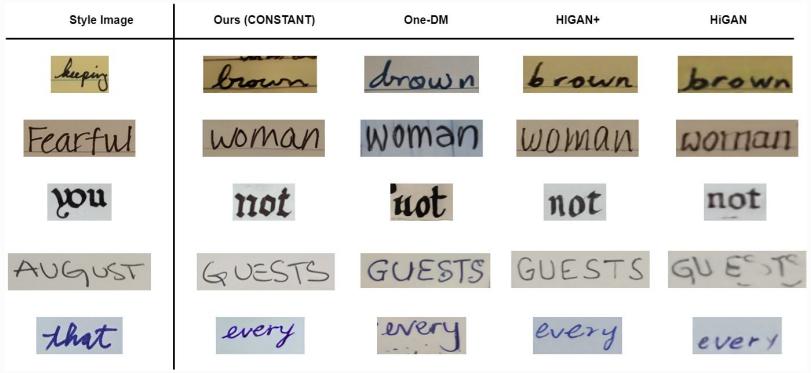
Multilingual Support

Outperforming SOTA by 10% on Chinese and Vietnamese structural generation.



IMGUR5K Diversity

Achieving superior 0.99 HWD on highly diverse, real-world data distributions.



Conclusion and Future work

Novel Framework: Introduced **CONSTANT**, a high-quality one-shot diffusion model for handwriting generation.

Key Innovations: Effectively captured intricate writer styles using **Style-Aware Quantization (SAQ)** and multi-scale **Patch Contrastive Enhancement**.

SOTA Performance: Achieved state-of-the-art results across English, Chinese, and Vietnamese datasets, outperforming both one-shot and few-shot competitors.

Robustness: Demonstrated superior generalization for **unseen styles** and **out-of-vocabulary** words..

Conclusion and Future work

Line-Level Generation: Expanding the model to generate longer, coherent lines of text rather than single words.

Artistic Styles: Enhancing style extraction for highly artistic or calligraphic handwriting where standard features are less distinct.

Dynamic Codebooks: Investigating adaptive codebook sizes to automatically match the stylistic complexity of different datasets.

Questions?

Thank you for your attention and interest in our research.

Anh-Duy Le | leanhduy497@gmail.com
Viettel Artificial Intelligence and Data Services Center