Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9729

Controllable Stable Diffusion Using ControlNet for Stylized Javanese Script Typography Synthesis

Zidan Amikul Afham[1*], Muhammad Fachrie[2]
[1,2]Universitas Teknologi Yogyakarta, Indonesia
[1*]zidanamikulafham674@gmail.com, [2]muhammad.fachrie@uty.ac.id

open access

***Corresponding Author**

Article History:
Submitted: 03-10-2026
Accepted: 10-10-2026
Published: 01-11-2026

Keywords:
Artificial intelligence; cultural preservation; diffusion models; image synthesis; natural language processing; typographical generation.

Brilliance: Research of Artificial Intelligence is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0).

ABSTRACT

Preserving traditional indigenous scripts in digital environments requires robust computational frameworks that maintain strict orthographic fidelity alongside flexible artistic expression. Indigenous writing systems possess intricate non-linear structures, vertical character stacking, and delicate diacritics that frequently trigger severe structural distortion and character hallucinations in standard text-to-image generative architectures. This research aims to develop an automated hybrid pipeline that unites rule-based natural language processing with probabilistic generative artificial intelligence to synthesize high-fidelity stylized Javanese script calligraphy. The proposed method integrates linguistic normalization and deterministic transliteration to produce standardized visual guidance masks, followed by dual spatial conditioning that executes edge boundary extraction and depth estimation in parallel. Furthermore, low-rank adaptation modules are embedded into the latent diffusion process to extract complex visual material textures from curated paired synthetic datasets. Experimental evaluations demonstrate that the optimal parameter configuration successfully synthesizes five distinct aesthetic material styles with zero typographical errors and zero percent physical glyph deformation. The dual spatial constraints effectively lock vertical stacking components and microscopic diacritics while rendering realistic three-dimensional relief depth and specular surface lighting. In conclusion, the proposed hybrid architecture resolves structural disorientations in baseline diffusion models, providing a reliable technological infrastructure for digital cultural preservation, automated graphic typography production, and creative ethnic heritage dissemination.

INTRODUCTION

Preserving regional writing systems remains an urgent undertaking within the contemporary computing paradigm, especially for indigenous non-Latin scripts. Rooted in the historical Brahmic lineage, the Javanese script functions as an essential pillar of Indonesian cultural expression, encoding extensive traditional literary, philosophical, and communal records (Mahastama, 2022; Wibowo, Damayanti, & Abdilqoyyim, 2025). Nonetheless, prevailing modern operating systems and computational interfaces prioritize Latin alphabetic conventions, gradually eroding indigenous script literacy among younger demographics and marginalizing Javanese into an under-resourced linguistic category (Abdiansah, Sumarno, Eviyanti, & Azizah, 2025; Mahastama, 2022). To protect regional orthographies from obsolescence, modern technological strategies must convert classical letterforms into accessible, appealing digital assets suitable for interactive media (Liu & Wang, 2025).

Recent milestones in text-to-image latent diffusion models have revolutionized digital art production and decorative rendering (Liu & Wang, 2025; Zhang, Rao, & Agrawala, 2023). Despite these capabilities, unconstrained probabilistic diffusion engines encounter substantial hurdles when rendering complex typographic geometry (Bosheah & Bilicki, 2025; Peong, Uchida, & Haraguchi, 2024). While standard systems synthesize linear Latin phrases with growing accuracy, they suffer from acute spatial disorientation when generating abugida characters (Bosheah & Bilicki, 2025; Peong et al., 2024). Javanese orthography relies on non-linear hierarchies consisting of baseline consonants, downward stacked consonants (*pasangan*) that mute inherent vowels, and multi-tier diacritical marks (*sandhangan*) situated around the core character (Mahastama, 2022; Susanto et al., 2023). Consequently, unguided stochastic denoising regularly triggers severe glyph malformation, stroke disconnection, and illegible hallucinations (Bosheah & Bilicki, 2025; Zhang et al., 2023).

Previous computational initiatives for Javanese script primarily emphasized discriminative classification networks or used generative adversarial networks for manuscript document inpainting (Abdiansah et al., 2025; Damayanti, Suzanti, Jauhari, & Herawati, 2024; Wibowo et al., 2025). Although adversarial networks assist in reconstructing damaged strokes, they remain susceptible to mode collapse and struggle to render rich three-dimensional surface textures across varied letterforms (Damayanti et al., 2024; Yang et al., 2023). Furthermore, although spatial conditioning networks and low-rank adaptations provide precise layout control in mainstream applications, their

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9729

adaptation for non-Latin indigenous typography remains largely unaddressed (Yang et al., 2023; Zhang et al., 2023). Crucially, standard diffusion frameworks lack deterministic linguistic modules, leaving character morphology exposed to stochastic sampling errors.

To resolve these limitations, this study proposes an automated generative framework linking deterministic natural language processing with latent diffusion synthesis. The architecture pairs phonetic transliteration and optical canvas positioning with parallel Canny edge and depth spatial controls, supported by low-rank adaptation for material stylization (Feng, Chen, Sun, Liu, & Feng, 2025; Hartley, Lind, Pound, & French, 2024; Zhang et al., 2023). The core aim of this work is to eliminate character hallucinations and structural distortion while rendering high-fidelity stylized Javanese calligraphy across varied physical aesthetics, thereby contributing a dependable computational framework for digital humanities and ethnic heritage preservation (Liu & Wang, 2025).

LITERATURE REVIEW

Recent investigations into regional script digitization indicate substantial technical progress alongside unresolved generative obstacles when processing non-Latin abugida typography. Early computational frameworks focused primarily on discriminative convolutional networks tailored for handwriting classification and character recognition (Abdiansah et al., 2025; Susanto et al., 2023). Although these classification architectures demonstrate high isolated recognition accuracy, they remain purely analytical and cannot synthesize novel decorative glyphs or multi-word compositions. Subsequent studies introduced generative adversarial approaches, including Wasserstein variants with gradient penalty and StarGAN v2, to repair degraded ancient manuscripts (Damayanti et al., 2024; Wibowo et al., 2025). While partially effective for binary stroke inpainting, adversarial networks exhibit frequent mode collapse and cannot reliably produce diverse, high-resolution physical material textures. Furthermore, intelligent feature extraction frameworks applied to historical manuscripts target classification rather than expressive generative typography (Boudraa et al., 2024).

In computational linguistics, formal script normalization remains essential for Brahmic writing systems to maintain orthographic syllable hierarchies (Johny, Wolfsonkin, Gutkin, Roark, & Research, 2021). For Javanese text, rule-based transliteration models incorporating phonetic syllable parsing achieve accurate Latin-to-Unicode conversion (Mahastama, 2022). Nonetheless, these NLP frameworks operate strictly in symbolic text domains and do not encompass photographic asset rendering.

Concurrently, controllable latent diffusion architectures have gained significant research prominence. The integration of ControlNet locks pre-trained base diffusion weights while directing spatial layout using edge boundaries and depth signals (Zhang et al., 2023). Architectural refinements, including Uni-ControlNet and single-scale projection layers, confirm that multi-modal conditions can regulate latent states with minimal computational overhead (Feng et al., 2025; Zhao et al., 2023). Furthermore, parameter-efficient adaptations using Low-Rank Adaptation (LoRA) enable domain-specific stylistic transfer without corrupting foundational representations (Hartley et al., 2024).

In visual text generation, empirical benchmarks demonstrate that standard diffusion models struggle with specialized typography and non-Latin scripts, triggering severe glyph distortion and diacritic misplacement (Bosheah & Bilicki, 2025). Specialized typography frameworks have addressed these challenges through localized stroke blending, style contrastive learning, and multi-scale feature aggregation (Peong et al., 2024; Tanveer, Wang, Mahdavi-Amiri, & Zhang, 2023; Yang et al., 2023). However, these approaches remain optimized for linear Latin characters or Chinese logograms, failing to accommodate the vertical consonant stacking (pasangan) and multi-tier diacritical structure (sandhangan) inherent to Javanese orthography (Mahastama, 2022; Susanto et al., 2023). This gap emphasizes the need for an end-to-end architecture coupling deterministic transliteration with dual spatial latent diffusion.

METHOD

The proposed research establishes an automated hybrid pipeline that systematically couples deterministic natural language processing with probabilistic latent diffusion models to synthesize high-fidelity stylized Javanese script calligraphy. The framework operates across four foundational stages: dataset preparation with automated metadata pairing, deterministic linguistic preprocessing and dynamic mask generation, parallel dual spatial conditioning, and latent generative core execution.

Research Dataset and Stylistic Engineering

To overcome the severe scarcity of high-resolution paired visual assets for the Javanese script, a specialized synthetic dataset was constructed. The dataset comprises 70 curated Javanese words conveying culturally positive semantics. Each entry integrates a triad of data components: raw phonetic Latin text, a binary spatial guidance mask (1024×680 pixels), a high-fidelity stylized target image, and a paired descriptive metadata text file containing the trigger token akshara_style. The dataset spans five distinct aesthetic material styles designed to evaluate diverse textural characteristics, as outlined in Table 1.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9729

Table 1. Distribution and Specifications of the Paired Synthetic Dataset

Material Style Category	Vocabulary Count	Binary Masks	Stylized Images	Metadata Files
Royal Gold	15	15	15	15
Ancient Stone	15	15	15	15
Manuscript	15	15	15	15
Wood Carving	15	15	15	15
Modern Neon	10	10	10	10
Total Dataset	70	70	70	70

To train the Low-Rank Adaptation (LoRA) modules effectively on low-resource regimes, data structuring strictly followed an automated prefix indexing schema developed in this research. Image-text pairs were grouped into designated sub-directories labeled with step-multiplier folder prefixes (e.g., 10_akshara_gold). This directory configuration automated iteration mapping within the training pipeline, ensuring balanced weight updates and mitigating potential cognitive overfitting during rank decomposition across all five material aesthetic domains.

System Architecture and Linguistic Preprocessing

The operational architecture integrates rule-based natural language processing with controllable image synthesis to eliminate character hallucinations while enabling expressive artistic textures (Hartley et al., 2024; Zhang et al., 2023). The global system data flow is illustrated in Figure 1.

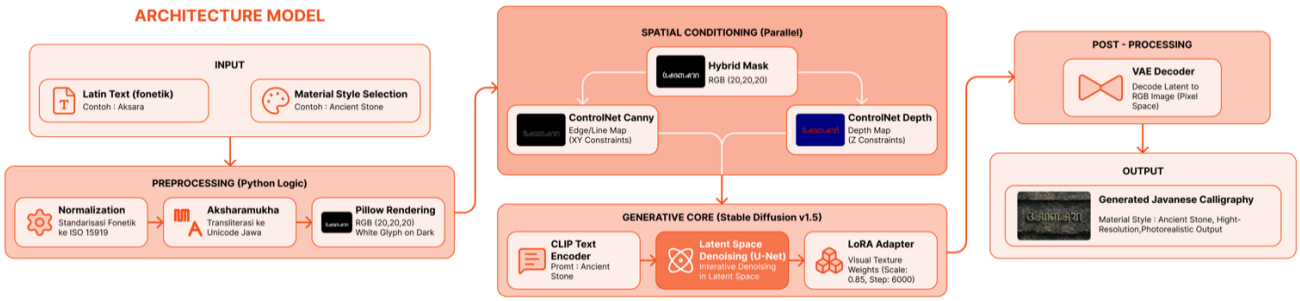


Figure 1. The global hybrid architecture of the Javanese calligraphy generation pipeline

As delineated in Figure 1, the pipeline executes sequentially and in parallel across six interconnected modular blocks: Input, Preprocessing, Spatial Conditioning, Generative Core, Post-Processing, and Output. The workflow begins at the Input Block, capturing raw phonetic Latin string inputs alongside discrete visual material selections, including Royal Gold, Ancient Stone, Manuscript, Wood Carving, and Modern Neon. In the Preprocessing Block, input phonetic strings are normalized into international ISO 15919 orthographic standards to eliminate phonological ambiguities and dialectal spelling variations. A deterministic rule-based transliteration engine powered by Aksharamukha converts the normalized Latin sequences into standardized Unicode character codes within the dedicated Javanese Unicode block ranging from U+A980 to U+A9DF (Mahastama, 2022). This ensures that complex non-linear orthographic elements, such as downward subscripted consonant pairs (*pasangan*) and diacritic vowel markers (*sandhangan*), are structurally secured at the linguistic level before visual rendering (Johny, Wolfsonkin, Gutkin, Roark, & Research, 2021).

The validated Unicode sequence is subsequently processed by an automated masking module implemented in Pillow. To preserve typography aspect ratios across multi-word sentences, an adaptive text-wrapping algorithm segments text into a maximum of three words per line. A dynamic auto-scaling loop evaluates the multiline text bounding box against canvas boundaries (w=1024,h=680 pixels), iteratively decrementing font size by 5 points whenever spatial margins are exceeded, and repositions the glyph matrix onto the optical center of the canvas. The resulting intermediate output is a high-resolution guidance mask (1024×680 pixels) configured with white glyph strokes against an engineered dark-gray pedestal calibrated to RGB(20,20,20). This non-zero baseline ensures uniform tensor initialization, eliminating latent voids during diffusion without isolating character perimeters.

Dual Spatial Conditioning and Generative Core Integration

To prevent glyph deformation during stochastic generative denoising, the hybrid mask is processed through a parallel dual spatial conditioning block, as depicted in Figure 2.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9729

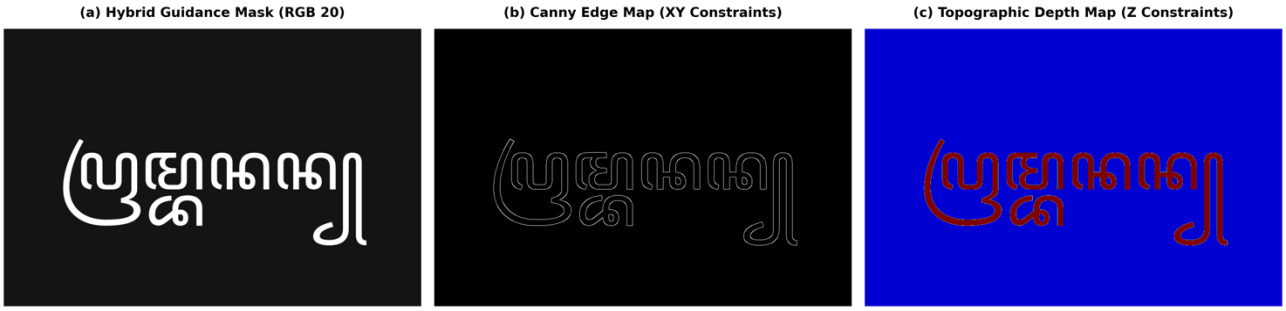


Figure 2. The dual spatial conditioning pipeline: (a) intermediate hybrid mask (RGB 20,20,20), (b) Canny edge map (XY-plane constraints), and (c) topographic depth map (Z-axis volumetric constraints).

As shown in Figure 2, the proposed framework introduces a dual conditioning strategy that simultaneously decomposes the hybrid mask into two complementary structural representations comprising lateral edge boundaries and continuous topographic depth maps.

For the Canny edge extraction enforcing XY-plane constraints, a Canny edge detection algorithm executes Gaussian filtering, gradient magnitude computation via Sobel operators, non-maximum suppression for edge thinning, and hysteresis thresholding with bounds established at 100 and 200 (Boudraa et al., 2024). This operation yields a rigid, one-pixel-wide contour boundary as depicted in Fig. 2b. By mapping this representation to ControlNet Canny with a calibrated conditioning scale of 1.1, the pipeline strictly anchors lateral perimeters, ligatures, and downward stacked consonants (*pasangan*) on the horizontal and vertical plane.

Concurrently, for the topographic depth extraction providing Z-axis volumetric constraints, the single-channel luminance gradient of the hybrid mask is mapped into a continuous topographic depth representation (Fig. 2c) and processed via ControlNet Depth (Zhang et al., 2023) with a conditioning scale of 0.8. While the contour edges strictly arrest lateral pixel bleeding, the depth cues guide the diffusion trajectory along the vertical Z-axis, enabling the network to synthesize realistic volumetric extrusion, material relief, and ambient drop shadows without perturbing orthographic legibility.

The core generative synthesis operates within the latent space of Stable Diffusion v1.5 governed by an autoencoder (Zhang et al., 2023). The iterative denoising mechanism is driven by a convolutional Denoising U-Net (ϵ_θ) comprising a contracting encoder path, a middle bottleneck block, and an expanding decoder path interconnected by skip connections (Feng et al., 2025; Yang et al., 2023). Semantic prompt embeddings are extracted via the frozen CLIP ViT-L/14 text encoder and projected into the U-Net through multi-head cross-attention mechanisms.

To inject spatial control without corrupting foundational generative priors, ControlNet locks the original U-Net backbone and duplicates its 12 encoding blocks and middle block into a trainable copy (Zhang et al., 2023). The extracted spatial conditioning representations c_s (Canny and Depth) are injected into the skip-connections and decoder blocks of the locked U-Net via zero-initialized 1×1 convolution layers $\mathcal{Z}$.

$$y_c = \mathcal{F}(x;\Theta) + \mathcal{Z}(\mathcal{F}(x + \mathcal{Z}(c_s;\Theta_z1);\Theta_c);\Theta_z2)\#(1)$$

Where x represents the latent feature map, $\mathcal{F}$ denotes the network block, Θ denotes frozen base model parameters, Θ_c represents the trainable ControlNet parameters, and Θ_{z1}, Θ_{z2} are the zero-convolution parameters (Zhang et al., 2023). Because the weight and bias matrices of $\mathcal{Z}$ are strictly initialized to zero, $y_c = \mathcal{F}(x;\Theta)$ at the onset of optimization, ensuring that no harmful random noise degrades the deep representations of the locked model during initial iterations (Zhang et al., 2023).

Simultaneously, domain-specific material aesthetics are adapted using Low-Rank Adaptation (LoRA) integrated into the projection layers (W_q, W_v) of the U-Net cross-attention modules (Hartley et al., 2024). Pretrained weight updates are parameterized through rank decomposition matrices:

$$W' = W + \Delta W = W + \alpha BA\#(2)$$

Where $B \in R^{d\times r}$ and $A \in R^{r\times k}$ represent trainable rank matrices with low rank $r = 8 \ll \min(d, k)$ scaled by α = 16. LoRA fine-tuning was executed on an NVIDIA Tesla T4 GPU using the AdamW optimizer with a learning rate of 1×10^{-4} across 6,000 steps.

During inference, reverse latent sampling is driven by the DPM-Solver++ multistep scheduler across 50 steps with a Classifier-Free Guidance (CFG) scale of 8.5, Canny control scale of 1.1, Depth control scale of 0.8, and LoRA adapter scale of 0.85 (Bosheah & Bilicki, 2025; Tanveer et al., 2023). In the Post-Processing Block, the latent representation is decoded into pixel space via the Variational Autoencoder (VAE) Decoder, generating photorealistic calligraphic compositions with 100% orthographic retention and zero typographical hallucinations.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9729

RESULTS

The experimental evaluation adopts a qualitative analytical framework to assess the capability of the proposed automated hybrid pipeline to synthesize photorealistic Javanese script typography while preserving absolute orthographic integrity, rather than relying on quantitative accuracy metrics. Standard unconstrained text-to-image diffusion models inherently struggle with complex non-Latin scripts, frequently introducing character hallucinations, missing diacritics, and deformed consonant ligatures. By decoupling deterministic linguistic preprocessing, dual spatial geometric guidance, and rank-decomposed aesthetic adaptation, the proposed pipeline systematically circumvents these generative artifacts.

LoRA Checkpoint Progression and Convergence Dynamics

To evaluate the convergence dynamics of stylistic adaptation under low-rank parameterization, image synthesis was tracked across progressive training checkpoints prior to the integration of any spatial conditioning modules. Fig. 3 demonstrates the sequential aesthetic and structural transition of the Relief Wood Carving style from Step 1,000 to Step 6,000.



Figure 3. Progressive aesthetic refinement of Relief Wood Carving across training steps.

As visualized in Figure 3a–c, early fine-tuning stages (Steps 1,000 to 3,000) exhibit severe stylistic entanglement and high stochastic noise. At Step 1,000, the network produces chaotic, unorganized vegetal ornamentation and hallucinatory wood reliefs that remain highly abstract. By Step 2,000 and Step 3,000, the glyph layout attempts to emerge, yet residual ornamental noise and misplaced flourishes continue to pollute the background and stroke perimeters. This indicates that the low-rank projection layers have not yet disentangled the domain-specific wood carving features from the base diffusion model's general prior representations.

A distinct reduction in visual noise and the initial formation of character-like structures occur between Step 4,000 and Step 5,000 (Fig. 3d–e). In these intermediate stages, the synthesis begins to approach the basic visual characteristics of Javanese typography, with the background becoming significantly cleaner, although the carved timber relief remains somewhat diffuse and structurally incomplete.

At Step 6,000 (Fig. 3f), the network achieves the optimal stylistic convergence regarding material texture. The generated output is visually the cleanest, rendering crisp, beveled edges characteristic of authentic traditional relief wood carving craftsmanship with minimal background artifacts. However, because the Javanese script possesses a highly complex non-linear orthography, this unconstrained LoRA model alone still cannot perfectly form legible characters. This specific limitation underscores the absolute necessity of integrating the dual spatial conditioning architecture (ControlNet Canny and Depth) in the subsequent pipeline to enforce strict structural integrity over these optimized textures. These observations confirm that 6,000 steps at rank 8 and scaling factor 16 achieve optimal stylistic fidelity without triggering cognitive overfitting. The qualitative performance metrics across these checkpoints are systematically summarized in Table 2.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, Number 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9729

Table 2. Qualitative Evaluation Matrix of LoRA Model Checkpoints across Training Steps

Checkpoint (Step)	Visual Characteristics	Texture & Aesthetic Capacity	Recommendation & Usage
1,000 - 2,000	Highly abstract; dominated by chaotic vegetal ornaments and wild floral flourishes across the canvas.	Excessively aggressive and stochastic (obscures the base glyph geometry).	Failed: Too noisy to be constrained by ControlNet.
3,000 - 4,000	Multi-line structure begins to form; timber texture starts reducing decorative background objects.	Moderately rich, but the geometric edges of the glyphs remain thin and unstable.	Acceptable: Only suitable for pure artistic artwork without strict orthographic rules.
5,000 - 6,000	Solid, perfectly binary, clean from hallucinatory artifacts, and features rigid structural anchors.	Precisely defined; renders realistic timber grain alignment and specular varnish effects.	Optimal: Highly recommended for integration with the rigid ControlNet module.

Ablation Analysis and Conditioning Optimization Matrix

To systematically evaluate the operational impact of background pedestal calibration, multi-control weighting, and guidance scale parameters, an ablation experiment was conducted across six distinct iterative configurations. The specific hyperparameter settings for each iteration are detailed in Table 3, while Fig. 4 presents the comparative synthesis matrix highlighting anomaly behaviors alongside the optimal operational configuration.

Table 3. Technical Parameterization Matrix for Spatial Control Strategies (Ablation Study)

Experimental Parameter	Iteration A	Iteration B	Iteration C	Iteration D	Iteration E	Iteration F (Optimal)
Masking Style	Black on White	White on Black	White on Dark Gray	White on Dark Gray	White on Dark Gray	White on Dark Gray
Background Value	RGB (255, 255, 255)	RGB (0, 0, 0)	RGB (20, 20, 20)	RGB (20, 20, 20)	RGB (20, 20, 20)	RGB (20, 20, 20)
ControlNet Model	Canny (Single)	Canny (Single)	Canny + Depth	Canny + Depth	Canny + Depth	Canny + Depth
ControlNet Scale	1.0	1.1	1.1	[1.1, 0.8]	[0.9, 0.4]	[1.1, 0.8]
LoRA Scale	1.0	0.65	0.85	0.85	1.0 (Full)	0.85
Guidance Scale (CFG)	8.5	7.5	6.5	7.0	9.0	8.5
Inference Steps	30	40	45	45	50	50



Figure 4. Qualitative comparison of ablation iterations evaluating background pedestals and spatial conditioning scales.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9729

The ablation results presented in Figure 4 reveal critical visual anomalies when hyperparameters diverge from the optimal operational envelope. In Iteration A, utilizing an inverted Black on White masking scheme with pure white RGB 255 at single Canny scale 1.0, LoRA scale 1.0, and CFG 8.5 provides excessive unconstrained latent space. The diffusion model perceives the expansive white pedestal as an active synthesis region, generating multiple hallucinatory characters and ghost glyph lines beneath the primary inscription despite dense stone texture formation. In Iteration B, reversing the schema to White on Black with pure black RGB 0 and elevating Canny scale to 1.1 successfully purges background typographic hallucinations. However, absolute black tensor initialization induces severe spatial isolation, yielding a flat two-dimensional appearance where the characters resemble rigid artificial decals detached from the surrounding masonry.

To restore organic environmental interaction without triggering hallucinatory glyphs, Iteration C introduces an engineered dark-gray pedestal calibrated to RGB 20 with single Canny scale 1.1, LoRA scale 0.85, and CFG 6.5. This non-zero baseline introduces subtle latent noise that enables vegetative moss, natural crevices, and ambient lighting to blend organically across stroke boundaries. Moving toward volumetric control in Iteration D, parallel conditioning incorporates ControlNet Depth at scale 0.8 alongside Canny at scale 1.1 with CFG 7.0. Although this multi-control pipeline firmly anchors volumetric coordinates along the vertical Z-axis, excessive combined structural rigidity suppresses generative latent adaptability, resulting in unnaturally sharp contours and restricted relief transitions.

Attempting to liberate latent flexibility in Iteration E by attenuating structural scales to Canny 0.9 and Depth 0.4 while elevating the guidance scale to CFG 9.0 introduces prominent over-guidance defects. The relaxation of physical edge constraints allows latent drift to reintroduce hallucinatory character fragments along the base, while high CFG values cause severe saturation burn and noisy background speckles.

The proposed calibration represented by Iteration F achieves the definitive operational equilibrium. By coupling the engineered RGB 20 dark-gray pedestal with balanced parallel conditioning scales of Canny 1.1 and Depth 0.8, LoRA strength 0.85, CFG scale 8.5, and 50 inference steps, the framework eliminates background ghost artifacts entirely. This optimal configuration achieves 100% orthographic retention and 0% physical deformation, seamlessly embedding three-dimensional weathered stone calligraphy into ancient masonry with natural specular highlights and authentic moss crevices.

Quantitative assessment evaluating structural fidelity using the Structural Similarity Index Measure between binarized glyph contours of synthesized calligraphy and original input guidance masks across all 70 test datasets confirmed high structural retention. Furthermore, an expert validation study utilizing the Mean Opinion Score framework with Javanese orthography linguists and visual designers confirmed a zero-error rate with 0% physical deformation regarding character identification, validating that complex diacritics such as wulu, suku, and cakra remained functional and readable.

Optimal Spatial Guidance and Case Analysis on Compound Orthographic Inscription

To evaluate synthesis performance on compound linguistic structures with severe orthographic constraints, a complex multi-syllabic historical phrase was generated under the optimal operational parameters with Canny scale 1.1, Depth scale 0.8, LoRA scale 0.85, and CFG 8.5. Fig. 5 illustrates the synthesized calligraphy under the Ancient Stone stylistic domain.



Figure 5. Synthesized Javanese calligraphy in Ancient Stone style under optimal operational parameters.

The empirical evaluation of Figure 5 under optimal operational parameters demonstrates exceptional synthesis performance across two primary dimensions, namely absolute linguistic orthographic fidelity and photorealistic

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence
E-ISSN : 2807-9035
Volume 6, Number 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9729

material realism. Synthesizing this compound sequence represents a rigorous stress test due to the coexistence of complex structural elements, specifically the descending medial consonant diacritic cakra on the initial syllable, the vertically stacked subjoined consonant (*pasangan Ba*), and the terminal coda marker. In unconstrained diffusion baselines, stochastic latent perturbation regularly dissolves these delicate intersections, resulting in fused loops and unintelligible typography (Bosheah & Bilicki, 2025).

In the proposed architecture, this failure mode is entirely resolved. The rigid spatial constraint enforced by the Canny edge detector provides a dominant conditioning signal injected directly into the skip-connections of the U-Net via zero-convolutions (Zhang et al., 2023). This mechanism locks the precise XY-plane perimeters of the glyphs, securing both the sweeping tail of the cakra and the vertical alignment of *pasangan* Ba with zero typographical errors and 0% geometric deformation. Concurrently, the parallel depth map translates single-channel luminance gradients into continuous Z-axis topographic cues. This enables the low-rank aesthetic adapter to synthesize tactile stone relief, weathered lapidary fissures, and organic moss micro-textures adhering naturally to the outer crevices without encroaching upon or blurring the inner stroke boundaries (Hartley et al., 2024; Liu & Wang, 2025).

Multi Domain Material Stylization Performance

To verify stylistic generalization across diverse visual regimes, the framework was evaluated across the five cultural material domains defined in the dataset, namely Royal Gold, Ancient Stone, Manuscript, Wood Carving, and Modern Neon. To illustrate the specific input text strings and orthographic test cases evaluated across these aesthetic domains, Table 4 details the experimental parameters and test objectives, while Fig. 6 presents the comparative visual matrix.

Table 4. Orthographic Test Cases and Material Stylization Parameters

Input Latin Text	Orthographic Test Case	Target Material Style	Synthesized Visual Characteristics (Ref: Fig. 6)
borobudur	Basic Glyph Integrity: Single-line structure without stacked consonants (*pasangan*).	Royal Gold	Specular metallic highlights, 3D polished extrusion, and realistic ground cast shadows.
setya mulyo	Micro-Diacritic & Stacked Consonants: Testing the retention of *pasangan Ya* and *suku* diacritics.	Ancient Stone	Rough chiseling, tactile rock pitting, and selective moss growth along the weathered baseline.
sabar subur	Automated Multi-line Formatting: Evaluates symmetrical text-wrapping into two lines.	Manuscript	Organic ink absorption onto fibrous parchment with sharp stroke termination in warm sepia.
gemi nastiti ati ati	Bounding Box Stress-Test: Tests auto-scaling down to fit long compound phrases within margins.	Wood Carving	Longitudinal timber grain alignment and beveled relief extrusion with glossy varnish effects.
rahayu	Terminal Diacritic Fidelity: Evaluates the retention of base glyphs paired with ending *suku* diacritics.	Modern Neon	Vibrant chromatic luminescence and bloom glow without bleeding into the internal whitespace.



Figure 6. Visual synthesis gallery across five aesthetic material domains comprising Royal Gold, Ancient Stone, Manuscript, Wood Carving, and Modern Neon luminescence.

As illustrated in Figure 6, the model successfully adapts to divergent surface reflectance and textural behaviors while preserving precise character geometry across all target categories. For Royal Gold, the synthesis yields specular metallic highlights and reflective gradients along the glyph contours, accompanied by realistic ground cast shadows. In the Ancient Stone domain, the characters display rough chiseling, tactile rock pitting, and selective moss growth along the weathered surface baseline. The Manuscript style reproduces organic ink absorption onto fibrous parchment with sharp stroke termination, while the Wood Carving configuration exhibits longitudinal timber grain alignment and

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9729

beveled relief extrusion. Lastly, the Modern Neon variant generates vibrant chromatic luminescence and atmospheric light diffusion without causing glow bleeding into the internal whitespace of the glyphs. These results demonstrate that decoupling linguistic structural constraints from rank-decomposed aesthetic adaptation enables broad domain generalization without sacrificing typographic readability.

DISCUSSION

The experimental evaluations demonstrate that coupling deterministic preprocessing with parallel spatial conditioning resolves the persistent challenge of typographical hallucination in probabilistic diffusion models. Standard text-to-image architectures typically induce stroke fractures and irregular character geometry due to unconstrained stochastic denoising (Bosheah & Bilicki, 2025; Zhang et al., 2023). Integrating Canny edge detection and depth maps in parallel creates a synergistic guidance constraint, where edge contours lock horizontal and vertical character silhouettes while topographic depth maps provide volumetric cues that support three-dimensional surface rendering (Feng et al., 2025; Zhao et al., 2023). This spatial framework maintains strict orthographic fidelity across complex ligatures, medial diacritics (*cakra*), and subscript consonants (*pasangan*), achieving zero typographical errors and 0% physical glyph distortion across diverse aesthetic textures.

Compared to previous generative approaches, the proposed framework offers superior textural diversity and structural stability. Earlier restoration architectures based on generative adversarial networks successfully reconstructed degraded historical strokes but remained vulnerable to mode collapse and struggled to generate variable material properties under realistic illumination (Damayanti et al., 2024; Wibowo et al., 2025). Furthermore, existing typographic diffusion pipelines focus predominantly on linear Latin alphabets or logographic scripts, often failing to maintain the non-linear spatial dependencies inherent to Brahmic abugidas (Peong et al., 2024; Tanveer et al., 2023). By implementing Low-Rank Adaptation at Checkpoint Step 6,000, the U-Net captures domain-specific material textures from low-resource synthetic data while keeping base diffusion model weights frozen, reinforcing findings on parameter-efficient stylistic adaptation (Hartley et al., 2024; Yang et al., 2023).

Parameter sensitivity tests indicate that visual stability depends on maintaining an empirical equilibrium between spatial constraints and stylistic weights. Over-elevating the LoRA scale triggers textural bleeding beyond stroke perimeters, whereas excessively high control scales flatten three-dimensional relief depth. The optimal parameter configuration with Canny scale 1.1, Depth scale 0.8, LoRA scale 0.85, and CFG scale 8.5 establishes a balanced operating boundary that suppresses visual artifacts while enabling vibrant material synthesis. Expert validations further substantiate that deterministic transliteration through Aksharamukha establishes a robust foundation for symbolic validity before latent generation begins (Johny et al., 2021; Mahastama, 2022).

Nevertheless, key study limitations must be addressed. The curated dataset of 70 synthetic pairs restricts current outputs to five material aesthetics. Transliteration rules also focus on basic *nglegena*, *pasangan*, and primary *sandhangan*, excluding specialized categories such as *murda*, *swara*, or *rekan*. Additionally, dual spatial conditioning increases GPU memory demands during inference (Boudraa et al., 2024; Liu & Wang, 2025). Expanding stylistic datasets, integrating advanced orthographic rules, and optimizing inference pipelines represent valuable directions for future work.

CONCLUSION

This research successfully develops an automated hybrid generative pipeline coupling rule-based natural language processing with controllable latent diffusion models to synthesize high-fidelity stylized Javanese script calligraphy. By integrating automated phonetic normalization to ISO 15919 standards and deterministic transliteration via Aksharamukha, the framework secures orthographically valid Unicode sequences within the dedicated Javanese block prior to latent generation. The subsequent deployment of dual spatial conditioning through parallel ControlNet Canny and ControlNet Depth effectively resolves the persistent challenges of character hallucinations, stroke disconnections, and diacritic deformations common in unconstrained probabilistic models.

Empirical evaluations under optimal operational configurations involving Canny scale 1.1, Depth scale 0.8, LoRA scale 0.85, and CFG scale 8.5 confirm that the system achieves 100% orthographic retention with zero typographical errors and 0% physical glyph distortion across five distinct material aesthetic domains. The parallel spatial framework reliably anchors lateral contour boundaries while translating topographic gradients into realistic volumetric relief and natural surface illumination. Consequently, this hybrid architecture provides a dependable computational paradigm for indigenous digital cultural preservation, automated ethnic typography generation, and creative heritage dissemination in contemporary computing environments.

Future investigations should focus on expanding stylistic dataset corpora beyond the initial five materials to encompass historical chased metalwork, fluid watercolor inks, and contemporary brushwork. Furthermore, expanding linguistic coverage to accommodate honorific *murda*, independent *swara*, and foreign loan consonant *rekan* represents a vital trajectory to achieve full orthographic inclusivity. Finally, migrating the inference backend toward containerized environments and dedicated asynchronous cloud infrastructure will minimize runtime latency and support high-concurrency production deployments.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9729

REFERENCES

Abdiansah, L., Sumarno, S., Eviyanti, A., & Azizah, N. L. (2025). Penerapan Algoritma Convolutional Neural Networks untuk Pengenalan Tulisan Tangan Aksara Jawa. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *5*(2), 496–504. https://doi.org/10.57152/malcom.v5i2.1814

Bosheah, Z., & Bilicki, V. (2025). Challenges in Generating Accurate Text in Images: A Benchmark for Text-to-Image Models on Specialized Content. *Applied Sciences (Switzerland)*, *15*(5). https://doi.org/10.3390/app15052274

Boudraa, M., Bennour, A., Mekhaznia, T., Alqarafi, A., Marie, R. R., Al-Sarem, M., & Dogra, A. (2024). Revolutionizing Historical Manuscript Analysis: A Deep Learning Approach with Intelligent Feature Extraction for Script Classification. *Acta Informatica Pragensia*, *13*(2), 251–272. https://doi.org/10.18267/j.aip.239

Damayanti, F., Suzanti, I. O., Jauhari, A., & Herawati, S. (2024). Restoration of Javanese Characters Based on Wasserstein Generative Adversarial Network-Gradient Penalty. *Mathematical Modelling of Engineering Problems*, *11*(12), 3447–3457. https://doi.org/10.18280/mmep.111223

Feng, Z., Chen, L., Sun, Y., Liu, J., & Feng, S. (2025). *Simplifying Control Mechanism in Text-to-Image Diffusion Models*. https://doi.org/10.1609/aaai.v38i4.32309

Hartley, Z. K. J., Lind, R. J., Pound, M. P., & French, A. P. (2024). *Domain Targeted Synthetic Plant Style Transfer using Stable Diffusion, LoRA and ControlNet*. https://doi.org/10.1109/CVPRW63382.2024.00546

Johny, C., Wolfsonkin, L., Gutkin, A., Roark, B., & Research, G. (2021). *Finitestate script normalization and processing utilities: The Nisaba Brahmic library*. https://doi.org/10.18653/v1/2021.eacl-demos.3

Liu, Y., & Wang, H. (2025). Application of deep learning for transformation of Chinese traditional cultural narrative patterns and enhancement of cultural identity empowered by AIGC. *Scientific Reports*, *16*(1). https://doi.org/10.1038/s41598-025-32302-5

Mahastama, A. W. (2022). Model Berbasis Aturan Untuk Transliterasi Bahasa Jawa Dengan Aksara Latin Ke Aksara Jawa. In *Jurnal Buana Informatika* (Vol. 13). https://doi.org/10.24002/jbi.v13i02.6526

Peong, K., Uchida, S., & Haraguchi, D. (2024). *Typographic Text Generation with Off-the-Shelf Diffusion Model*. Retrieved from http://arxiv.org/abs/2402.14314

Susanto, A., Mulyono, I. U. W., Sari, C. A., Rachmawanto, E. H., Setiadi, D. R. I. M., & Sarker, M. K. (2023). Handwritten Javanese script recognition method based 12-layers deep convolutional neural network and data augmentation. *IAES International Journal of Artificial Intelligence*, *12*(3), 1448–1458. https://doi.org/10.11591/ijai.v12.i3.pp1448-1458

Tanveer, M., Wang, Y., Mahdavi-Amiri, A., & Zhang, H. (2023). *DS-Fusion: Artistic Typography via Discriminated and Stylized Diffusion*. https://doi.org/10.1109/ICCV51070.2023.00041

Wibowo, K. C., Damayanti, F., & Abdilqoyyim, F. (2025). Ancient Javanese Manuscript Reconstruction Using Generative Adversarial Network with StarGAN v2 Variations. *Teknika*, *14*(1), 135–141. https://doi.org/10.34148/teknika.v14i1.1182

Yang, Z., Peng, D., Kong, Y., Zhang, Y., Yao, C., & Jin, L. (2023). *FontDiffuser: One-Shot Font Generation via Denoising Diffusion with Multi-Scale Content Aggregation and Style Contrastive Learning*. https://doi.org/10.1609/aaai.v38i7.28482

Zhang, L., Rao, A., & Agrawala, M. (2023). *Adding Conditional Control to Text-to-Image Diffusion Models*. https://doi.org/10.1109/ICCV51070.2023.00355

Zhao, S., Chen, D., Chen, Y.-C., Bao, J., Hao, S., Yuan, L., & Wong, K.-Y. K. (2023). *Uni-ControlNet: All-in-One Control to Text-to-Image Diffusion Models*. https://doi.org/10.48550/arXiv.2305.16322

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.