Apollo Journals
Engineering and Computing Insights
ECI
Vol. 01 Issue 01 2026

Review

From Generation to Evaluation: A Review of Quality Assessment Methods and Development Trends for Artificial Intelligence-Generated Fonts

Yuxin Li

School of Computer Science; Tongji University, Shanghai 200000, China; Email: 2452088@tongji.edu.cn

Abstract: Generative artificial intelligence (GAI) has advanced font generation toward greater diversity, personalization, and automation. However, assessing generated fonts remains challenging because quality depends on character identity, glyph structure, style, readability, and aesthetic preference. This structured narrative review synthesizes literature on font generation and quality assessment using four categories: traditional image-quality metrics; character correctness and structural assessment; style, readability, and aesthetic assessment; and multimodal or large-model-based evaluation. The review distinguishes direct typography-related evidence from methods transferred from general image-quality and multimodal research. Across the literature examined, pixel-level and distribution-level metrics remain useful baselines but cannot alone establish character correctness or typographic suitability. Structure-aware methods and human-centred assessment address complementary dimensions, while vision-language and large-model approaches remain promising but require task-specific validation. The review proposes an integrative generation–evaluation–diagnosis–optimization framework and identifies priorities for transparent benchmarks, interpretable sub-scores, multilingual testing, and reproducible human evaluation.

Keywords: AI-Generated Fonts, Font Quality Assessment, Generative Artificial Intelligence, Multimodal Large Models, Structure-Aware Evaluation, Aesthetic Evaluation

Received: 22 September 2026
Revised: 31 October 2026
Accepted: 18 November 2026
Published: 23 November 2026

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

1.1. Background and Significance

In recent years, the rapid development of generative artificial intelligence (GAI) has moved artificial intelligence-generated content (AIGC) into a new stage. New-generation generative models represented by generative adversarial networks (GANs), diffusion models, and multimodal large models have achieved major advances in image generation, artistic design, digital media, and cultural creativity [1][2][3]. Font generation is an important application of AIGC. Given a small number of samples or a textual description, font-generation systems can automatically produce new characters in a specified style, creating broad opportunities for digital typeface design, cultural-heritage preservation, personalized typography, human-computer interaction, and digital art creation[4].

Compared with natural images, font images are subject to much stricter structural and semantic constraints. A font must preserve the intended character content while satisfying requirements concerning stroke connections, component arrangement, center-of-gravity balance, and overall stylistic consistency. This issue is particularly important for Chinese fonts, in which an individual character may contain complex hierarchical structures and spatial relationships. Even when the overall visual appearance is satisfactory, missing strokes, misplaced components, or imbalanced proportions can seriously impair readability and aesthetics. Recent systems based on models such as Stable Diffusion and FontCLIP demonstrate strong artistic expressiveness, yet generated fonts may still exhibit broken strokes, structural errors, style drift, and semantic distortions[4][5].

Consequently, human subjective evaluation alone is insufficient for the rapid iteration and large-scale deployment of font-generation models. A scientific, unified, and automated quality-assessment framework has therefore become an important component of font-generation research. A high-quality evaluation system can not only objectively compare the generation capabilities of different models, but can also provide optimization targets for model training and promote font-generation technology toward greater stability, interpretability, and intelligence.

1.2. Specificity of Font Quality Assessment and Limitations of Existing Research

Most existing font-generation studies continue to adopt image-quality assessment methods developed for natural images, including peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), Fréchet Inception Distance (FID), and Learned Perceptual Image Patch Similarity (LPIPS) [6][7][8]. These metrics mainly focus on overall pixel consistency or perceptual similarity. Font evaluation, however, goes far beyond visual similarity because font images are simultaneously constrained by character content, glyph structure, and artistic style.

First, fonts have strict character-content constraints. An evaluation system must ensure semantic consistency between a generated character and its target character and avoid incorrect, missing, or structurally distorted characters. Second, fonts have explicit structural constraints, including stroke completeness, component relationships, topological connections, and spatial layout. These fine-grained features often determine whether a glyph conforms to writing conventions, yet conventional image metrics have difficulty detecting local structural defects. Third, fonts possess distinctive stylistic and aesthetic constraints. Typeface families differ substantially in stroke width, terminal and brush characteristics, rhythmic patterns, and overall artistic expression. Human judgments integrate visual perception with aesthetic experience, whereas traditional metrics lack effective models of stylistic consistency and aesthetic preference.

Researchers have recently applied vision-language models and multimodal large models to typography-related representation and evaluation. FontCLIP is a typography-specific vision-language model designed for semantic font representation, attribute prediction, and retrieval; it should not be described as a dedicated font-generation model. For general AI-generated-content (AIGC) image-quality assessment, methods such as patches sampling-based contrastive regression (PSCR) and text-image encoder-based regression (TIER) combine image features with semantic information, but their results do not by themselves validate them for font-quality assessment [9][10]. The literature reviewed here does not establish a single standardized benchmark that jointly measures character correctness, structure, style, readability, and aesthetics; this is a conclusion about the reviewed evidence, not a claim that no relevant resource exists anywhere.

1.3. Scope and Contributions

This paper is a structured narrative review focused on the evaluation of AI-generated fonts, rather than a systematic review following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. It contributes in three ways. First, it organizes evaluation methods into four categories and compares what they measure, their reference-data requirements, representation levels, and limitations. Second, it separates direct font- and glyph-related evidence from transferable methods developed for general image-quality assessment, font retrieval, calligraphy generation, and visual-text generation. Third, it synthesizes these strands into a proposed generation–evaluation–diagnosis–optimization framework, with particular attention to multilingual scripts and human-centred assessment. This scope differs from the broader survey by Wang et al. [5], which primarily reviews deep-learning methods for Chinese font generation: the present review focuses on how generated-font quality can be evaluated and how complementary evaluation signals can be combined. The proposed framework is an integrative synthesis by the authors, not an already standardized framework in the field.

In addition, the gap between generation and evaluation has become more pronounced as models increasingly generate visually plausible fonts. A useful evaluation system must distinguish between errors that are perceptually minor and errors that alter character identity or writing structure. For example, a small pixel deviation may be acceptable when a style intentionally changes, whereas a missing stroke may fundamentally change a Chinese character. This distinction explains why evaluation metrics designed for natural-image reconstruction cannot simply be transferred to typography without modification.

The significance of this problem is both methodological and practical. Methodologically, a unified assessment framework can provide common criteria for comparing generation models and diagnosing their failure modes. Practically, reliable evaluation can support model training, automatic quality control, font recommendation, intelligent design assistance, and large-scale production. The literature therefore needs to be considered not only as a collection of individual metrics but also as an emerging evaluation ecosystem in which image-level, structure-level, semantic, and human-centered evidence are jointly interpreted.

2. Literature Review

2.1. Development of AI Font Generation

Font generation has evolved from early pixel-level style transfer to increasingly structure-aware, diffusion-based, and semantically controlled generation. GAN-based methods established the foundation for few-shot font generation, while subsequent work introduced memory mechanisms, glyph-aware modules, component sharing, and transformer-based global structure modeling. More recently, diffusion models and multimodal foundation models have expanded the ability to generate complex glyphs and control style through textual or visual semantics. This progression has also exposed a fundamental limitation of conventional evaluation: as

https://doi.org/10.67728/eci.2026.004

generated fonts become more visually realistic, pixel-level similarity becomes less capable of representing structural correctness, semantic fidelity, stylistic consistency, and human aesthetic experience.

2.1.1. GAN-Based Font Generation

Goodfellow et al. [1] introduced GANs, in which a generator and discriminator are trained adversarially to produce realistic images. Researchers subsequently applied GANs to font generation with the goal of generating large character sets in a target style from a small number of examples. Early methods borrowed the image-to-image translation paradigm and mapped standard fonts to target fonts. Azadi et al.[11] proposed Multi-Content GAN for few-shot font style transfer, separating content and style encodings to learn their disentangled representations. MX-Font uses multiple localized experts to represent diverse local style patterns and improve generalization to unseen components and scripts [12]. DG-Font instead uses deformable generative networks for unsupervised font generation [13]. These methods broaden the few-shot generation landscape, although generation performance should not be conflated with a dedicated font-quality evaluation framework. Cha et al. [14] proposed DM-Font, using memory augmentation to increase knowledge sharing among characters and improve the generation of complex Chinese components and unseen characters. Gao et al. [15] proposed AGIS-Net, which jointly synthesizes glyph shape and texture from a small number of stylized examples using separate content/style encoders and collaborative decoders. These methods highlight the importance of explicit glyph information in font generation.

In evaluation, GAN-based font-generation studies commonly employ L1 or L2 loss, SSIM, and PSNR to measure differences between generated and reference fonts, while later studies introduced perceptual metrics such as FID and LPIPS. Nevertheless, GAN training can suffer from mode collapse and instability, especially for structurally complex Chinese fonts. End-to-end pixel mapping also lacks explicit modeling of strokes, radicals, components, and topological relationships, which can lead to stroke adhesion and missing components.

2.1.2. Structure-and Component-Aware Font Generation

Researchers gradually recognized that pixel-level mappings alone cannot adequately express the complex geometric structure inside fonts. Chinese characters consist of multiple radicals and components, numerous strokes, and intricate spatial arrangements under strict topological constraints. Structure-aware approaches therefore extend font representations from pixels to multilevel representations such as strokes, skeletons, components, and complete glyphs. DM-Font incorporates character-structure information to establish knowledge sharing among fonts and learn common structural features across characters [14]. VQ-Font combines vector quantization with structure-enhancement modules and learns font styles in a discrete latent space, reducing structural distortion in complex characters and improving consistency in style transfer.

Component-aware methods place greater emphasis on shared relationships among Chinese radicals and components. Park et al. [16] proposed CCFont, which decomposes Chinese characters into components and recombines them to generate fonts, including characters with complex structures. Chen et al. [17] introduced a transformer-based font-generation model that uses self-attention to learn long-range dependencies among strokes, components, and complete characters while preserving local details and global layout. Compared with convolutional networks, such approaches provide sFutronger global structure modeling capabilities.

2.1.3. Diffusion Models and Multimodal Large Models

Diffusion models generate samples by progressively adding noise and learning the reverse denoising process. They can model data distributions more stably and offer advantages in complex structural modeling, detail restoration, and generation diversity. Latent diffusion models reduce computational costs while maintaining generation quality [3]. Fu et al. [18] proposed MSD-Font, a multi-stage diffusion font framework inspired by professional font-design workflows. It divides generation into structural construction, font transfer, and detail refinement, using two networks to coordinate diffusion generation. The method improves stroke completeness and local detail quality for complex Chinese characters and reduces stroke adhesion and missing components.

Yang et al. [19] proposed FontDiffuser for one-shot font generation, combining multiscale content aggregation with style contrastive learning. Its results indicate that diffusion-based generation can outperform earlier GAN-based approaches while preserving character structure and enhancing style expression. Foundation models and multimodal large models have further shifted font generation from image generation toward semantically driven design. FontCLIP fine-tunes a pretrained Contrastive Language–Image Pre-training (CLIP) model with font-specific knowledge and constructs a unified semantic representation space for multilingual font retrieval, semantic style matching, and font-attribute understanding[4]. Tuo et al. [20] proposed AnyText for multilingual visual text generation and editing, illustrating the movement toward controllable, semantically driven visual-text generation. Overall, these developments increase the need for evaluation methods capable of jointly assessing structure, style, semantic control, and human perception.

2.2. Development of Font Quality Assessment

https://doi.org/10.67728/eci.2026.004

Font quality assessment has evolved in parallel with generation technology. Early work primarily adopted full-reference image metrics and distribution-based metrics. Later research introduced character recognition, skeleton and topology analysis, component-level assessment, local-region analysis, style and readability evaluation, aesthetic judgment, and multimodal semantic evaluation. This progression indicates a shift from measuring image similarity toward evaluating whether a generated font is semantically correct, structurally complete, stylistically coherent, readable, aesthetically acceptable, and suitable for real-world use.

Traditional metrics such as PSNR and SSIM are computationally simple and useful when a precise reference image is available, while Fréchet Inception Distance (FID) and Kernel Inception Distance (KID) evaluate differences between generated and real distributions. However, these metrics operate mainly at the image or feature-distribution level. OCR accuracy directly measures semantic recognizability; skeleton and topology analysis captures stroke connections and geometric structure; component-level measures can identify missing or displaced radicals; and local-region methods increase sensitivity to fine defects. Style evaluation has increasingly incorporated vision-language representations, while readability assessment uses reading speed, recognition accuracy, reading errors, and OCR performance. Aesthetic evaluation generally relies on expert ratings, user studies, mean opinion score (MOS), or Likert scales, reflecting the inherently subjective nature of visual preference.

Multimodal evaluation represents a further development. FontCLIP aligns font images and textual descriptions in a shared semantic space, enabling style-attribute prediction, multilingual retrieval, and semantic similarity assessment[4]. Large language models can potentially combine OCR outputs, structural analyses, and style descriptions to produce interpretable evaluation reports. The emerging paradigm of using large language models as judges (LLM-as-a-Judge) has already been studied extensively in language and multimodal generation [21] [22]. However, vision-language models may recognize textual content without fully understanding typography, a phenomenon described as the "Typography Gap" [23]. This suggests that future font assessment must combine multimodal models with domain-specific typography knowledge.

3. Materials and Methods

3.1. Literature Search Strategy

This paper uses a structured narrative-review approach to examine research on AI-generated font quality assessment. The initial manuscript identified IEEE Xplore, Web of Science Core Collection, and Google Scholar as relevant sources, and the review draws on targeted searches and citation tracing across computer vision, artificial intelligence, computer graphics, typography, and related fields. However, complete database-by-database query logs, exact search-run dates, and stage-wise record counts were not retained in the working records. The available evidence therefore does not support a reproducible PRISMA flow diagram or reliable numerical counts of identified, screened, excluded, and included records. To avoid implying a level of procedural completeness that cannot be demonstrated, the paper is explicitly framed as a structured narrative review and does not report invented screening counts.

Search concepts covered four topic groups: (1) AI-based font generation and font synthesis; (2) font quality assessment and glyph evaluation; (3) character structure, correctness, legibility, and readability; and (4) general image-quality assessment and AIGC evaluation methods potentially transferable to typography. Representative terms included "font generation," "font synthesis," "font quality assessment," "glyph quality," "glyph structure," "character correctness," "font readability," "image quality assessment," and "AIGC image quality assessment." These are representative concepts, not a claim that a complete reproducible Boolean string was preserved for every database.

Foundational work published before 2019 was identified through backward citation tracing and its role was to explain concepts such as adversarial generation, image-quality metrics, and perceptual assessment. Such publications, along with review articles, are treated as background or contextual sources rather than as part of a formally counted set of primary studies. The review does not claim an exhaustive census of all publications in the field.

The literature synthesis was informed by 36 cited publications in total, including primary studies, review articles, and foundational methodological works. Tables 1–3 present representative studies and resources selected from this body of literature to illustrate the principal methodological approaches and research directions.

3.2. Eligibility and Evidence Classification

The synthesis prioritizes primary studies that propose, apply, or experimentally examine methods relevant to font or glyph generation and assessment. Relevant tasks include character correctness, structural integrity, style consistency, legibility, readability, aesthetic quality, and evaluation methods that can be examined in a typography-related task. Studies on font generation are included where they clarify common quality requirements or failure modes, even when they do not propose a dedicated evaluation metric.

Studies focused exclusively on artistic design without a relevant technical or evaluation component, or providing insufficient information to assess their relevance, were not used as substantive evidence in the synthesis. Duplicate versions of the same work were treated as one publication record, with the peer-reviewed version preferred when available. Review articles were used to

establish context and identify primary work, but were not treated as primary empirical evidence. Because the source records do not preserve complete screening counts, the review does not claim a fully reproducible systematic-screening process.

To avoid conflating different research tasks, the literature is discussed in three evidence groups. Direct font-related evidence comprises work on font or glyph generation, font-specific evaluation, or typography-related assessment. Transferable methodological evidence comprises general image-quality, vision-language, or large-model evaluation methods that may inform font assessment but have not necessarily been validated on font data. Background evidence comprises foundational methods and review articles used to explain concepts and research context.

This distinction is applied throughout the Results: a method developed for general AIGC images is not treated as a validated font-evaluation method unless typography-specific evidence supports that interpretation. Similarly, font retrieval, calligraphy generation, and visual-text rendering are discussed only for the aspects relevant to evaluation; they are not assumed to be equivalent to font-family generation.

3.3. Analytical Framework

The literature was analyzed along two complementary axes: what aspect of quality is evaluated and how that quality is evaluated.

The first axis concerns what is evaluated: (1) character correctness, or whether the glyph preserves the intended character identity; (2) structural integrity, including stroke completeness, component relationships, and spatial organization; (3) style and semantic consistency, including correspondence to an intended style or description; and (4) human-centred visual quality, including legibility, readability, and aesthetic preference.

The second axis concerns how quality is evaluated: pixel-level and perceptual image metrics, distribution-based metrics, character-recognition methods, skeleton and topology analysis, component- and region-level analysis, human-subject assessment, vision-language models, and emerging large-model-based evaluation. These method categories overlap: one method may contribute evidence to several quality dimensions, and several methods may assess the same dimension from different perspectives.

The comparison tables in Section 4 map representative methods and studies to the analytical variables available in the cited publications: task, generation paradigm where applicable, evaluation target, reference-data requirement, representation level, script or language when stated, human involvement, and reported limitations. These tables are representative rather than exhaustive because the review does not have a complete, auditable study-level screening log. Proposed component-level indicators and emerging model-based judgments are explicitly labelled as potential or requiring validation when they are not established and validated metrics in the cited work.

The review considers font-generation research to identify quality requirements and common failure modes. Work on calligraphy generation, visual-text generation, font retrieval, and general AIGC assessment is included only where it provides contextual or transferable evidence. FontCLIP is treated as a typography-specific semantic representation and retrieval model [4], whereas AnyText is treated as a multilingual visual-text generation and editing method, not a font-family-generation model [20].

Table 1. Representative Study-Characteristics Map Based on the Cited Works.

Representative Work	Primary Task / Role	Paradigm or Method	Evaluation Target / Evidence	Script / Language	Evidence Group and Caveat
Azadi et al.	Few-shot font style transfer	GAN-based generation	Generation quality; image-level comparison in generation setting	Font datasets; script depends on study setup	Direct font-related generation evidence; not a dedicated evaluation framework
Cha et al.	Few-shot compositional font generation	Dual-memory / component-aware generation	Glyph fidelity and style transfer	Korean and Thai in the cited ECCV work	Direct font-related evidence; use the official 2020 publication
Park et al.	Few-shot font generation	Localized style representations and factorization	Content/style generation quality	Font-generation datasets	Direct font-related evidence; generation metrics should not be equated with a universal quality score
Yang et al.	One-shot font generation	Diffusion with multi-scale content aggregation and style contrastive learning	Generated glyph quality and style/content consistency	Font-generation dataset	Direct font-related evidence; published method is a generator, not a standalone evaluator
Tatsukawa et al.	Typography representation / retrieval	Font-specific vision-language model (FontCLIP)	Style attributes, semantic representation, retrieval	Multilingual font applications	Direct typography-related evidence; not a general all-dimensions quality score

https://doi.org/10.67728/eci.2026.004

Representative Work	Primary Task / Role	Paradigm or Method	Evaluation Target / Evidence	Script / Language	Evidence Group and Caveat
Yuan et al.	General AIGC image-quality assessment	PSCR / TIER	Image quality and text-image features	General generated-image data	Transferable methodological evidence; font-specific validity is not established by these works alone
Legge et al.	Font visual accessibility	Behavioral measures and optical character recognition (OCR)-based estimates	Reading acuity, critical print size, reading speed, preference; OCR proxy	Latin typefaces in the reported study	Typography/accessibility evidence; publication corrected to PLOS ONE (2026)
AnyText	Multilingual visual-text generation/editing	Text-conditioned image generation/editing	Rendered text within images	Multilingual visual text	Contextual evidence only; not font-family generation

Note: This is an illustrative evidence map, not a complete included-study census, because the original screening log and counts are unavailable. Source: Author.

4. Results

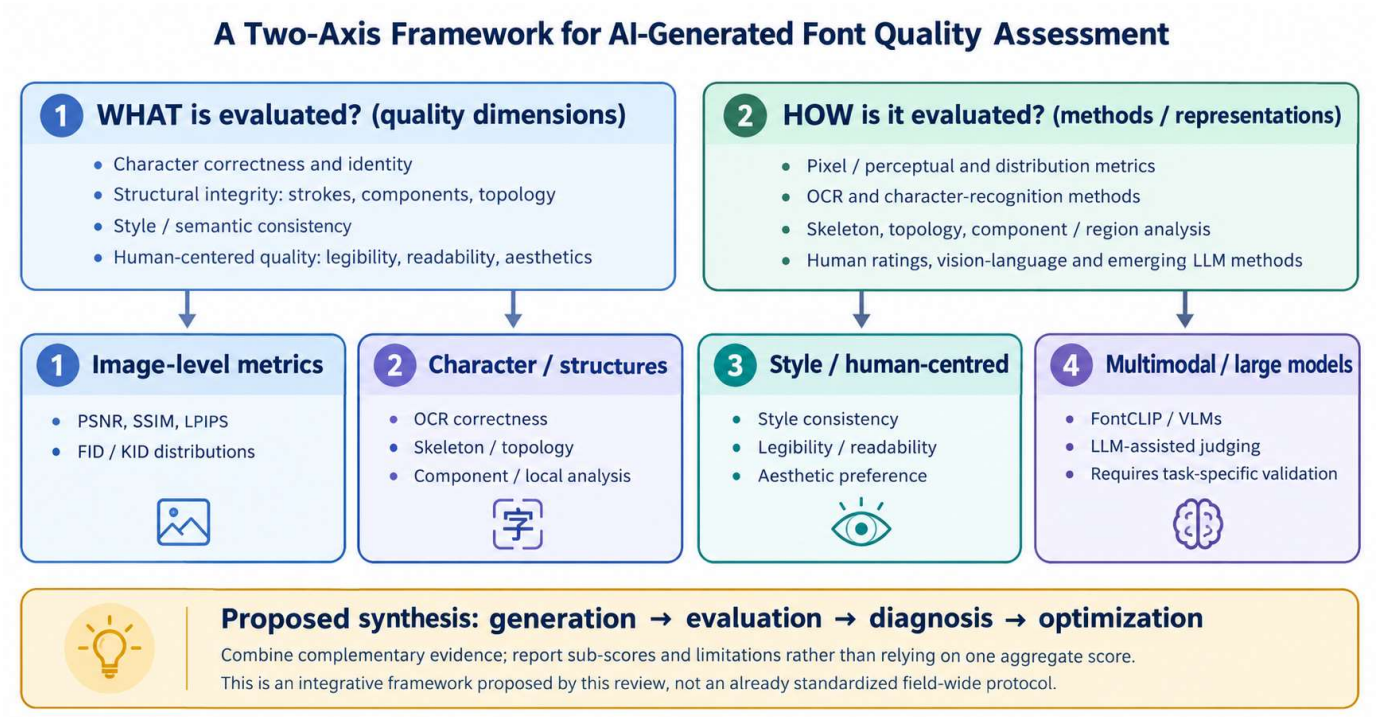

Figure 1. Two-Axis Taxonomy of Font-Quality Assessment and the Proposed Generation–Evaluation–Diagnosis–Optimization Synthesis. Note: The four categories organize evaluation methods; quality dimensions describe what is assessed. The lower pipeline is the authors' integrative proposal, not an established standard. Source: Author.

4.1. Traditional Image-Quality Metrics

Early AI-generated font research commonly adopted image-quality metrics developed for general image reconstruction and generation. These metrics provide useful baselines, but their ability to represent font-specific quality depends on the evaluation task and the availability of reference data.

Full-reference metrics compare a generated glyph with a corresponding reference image. Peak signal-to-noise ratio (PSNR) measures pixel-level reconstruction error and is computationally simple, but is sensitive to pixel misalignment and does not directly represent typographic quality[24]. Structural similarity (SSIM) compares luminance, contrast, and structural information and may provide a more perceptually meaningful comparison than PSNR in some settings[7]. Both can penalize legitimate stylistic variation when a generated glyph differs from a reference while preserving character identity and design intent.

Learned perceptual metrics, such as the learned perceptual image patch similarity (LPIPS), compare deep feature representations rather than relying solely on pixel differences[8]. Such representations may better reflect some perceptual similarities, but their

effectiveness for font assessment depends on whether the feature space captures typographic structure and style. A metric developed for natural images should therefore not automatically be interpreted as a validated measure of font quality.

Distribution-based metrics evaluate generated samples at the dataset level. Fréchet Inception Distance (FID) compares feature distributions of real and generated images, while Kernel Inception Distance (KID) uses a maximum mean discrepancy (MMD)-based approach to compare feature distributions [6][25]. These metrics describe distributional similarity but do not establish whether individual characters are correct, structurally complete, or readable. Their interpretation depends on the suitability of the feature representation and the composition of the evaluation dataset.

Overall, traditional image-quality metrics are useful for reproducible baseline comparisons, especially when reference images or comparable sample distributions are available. Nevertheless, a high score does not necessarily imply that a generated character is correct or that a font is suitable for practical use. These limitations motivate the use of complementary structural, semantic, and human-centered evaluation.

Table 2. Qualitative Comparison of Representative Font-Quality Evaluation Methods.

Metric / Method	What It Measures	Reference Required	Representation Level	Strengths	Limitations / Interpretation
PSNR	Pixel-wise reconstruction error	Yes, paired reference image	Pixel	Simple, inexpensive, reproducible baseline	Sensitive to alignment and style variation; does not establish glyph correctness
SSIM	Luminance, contrast, local structural similarity	Yes, paired reference image	Local image structure	More structure-aware than raw pixel error in some settings	Can miss semantic character errors; sensitive to reference choice
LPIPS	Distance between deep visual features	Usually paired images	Perceptual feature	Can capture perceptual differences not reflected by pixel error	Feature space is not typography-specific unless validated
FID	Distance between real/generated feature distributions	Reference set / real-sample distribution	Dataset / feature distribution	Summarizes broad distributional differences	Not a per-glyph score; does not directly measure character identity
KID	MMD-based distance between feature distributions	Reference set / real-sample distribution	Dataset / feature distribution	Distribution comparison with an estimator based on feature samples	Depends on feature representation and dataset composition
OCR recognition accuracy	Whether recognized content matches target glyph/text	Target character or transcription required	Character / semantic content	Directly tests content recognition	Recognizer bias and training data affect results; not equivalent to human readability
Skeleton / topology analysis	Stroke connectivity, skeleton structure, topological defects	Often a reference skeleton or explicit rules	Structure / topology	Can localize structural failures invisible to global metrics	Skeleton extraction can be noisy; scripts and stroke styles vary
Component-level indicators*	Radical/component presence, placement, or local matching	Typically requires annotations, reference components, or explicit rules	Component / local region	Potentially diagnoses missing or displaced components	Not one standardized metric; define formulas and validate before treating as established
Human ratings / mean opinion score (MOS)	Perceived quality, preference, or task-specific judgments	No image reference necessarily; rubric and samples required	Human / perceptual	Captures contextual and aesthetic dimensions	Costly; affected by rater population, instructions, and inter-rater variation
FontCLIP / vision-language similarity	Semantic font attributes and image-text or font-text alignment	Text prompts/labels or comparison items; depends on task	Multimodal semantic representation	Useful for style semantics and retrieval	Not a complete font-quality metric; validate against typography tasks
large language model/vision-language model (LLM/VLM)-assisted judging	Model-generated assessment or explanations from images/text	Depends on prompt and scoring design	Multimodal / reasoning-based	May combine signals and produce interpretable reports	Emerging for font assessment; prompt sensitivity, bias, and human agreement require validation

https://doi.org/10.67728/eci.2026.004

Note: Component-level indicators are potential operationalizations synthesized in this review, not claimed as a single established standardized metric. Source: Author.

4.2. Character Correctness and Structural Assessment

Unlike natural images, font images represent discrete linguistic symbols whose identity and internal structure must be preserved. A glyph can appear visually plausible while containing an incorrect character, a missing stroke, a malformed component, or an inappropriate connection. Character correctness and structural integrity should therefore be evaluated separately from global image similarity.

4.2.1. Character Recognition and Content Correctness

Optical character recognition (OCR) can be used to determine whether a generated glyph is recognized as its intended character. When the target character labels are known, recognition accuracy and confusion matrices can reveal content errors and systematic confusions between visually similar characters. This provides evidence about semantic recognizability and character-content fidelity.

However, OCR accuracy should not be interpreted as a complete measure of font quality. Recognition performance depends on the OCR model, image resolution, preprocessing, and script coverage. A glyph may be recognized correctly despite noticeable stylistic defects, while a highly stylized but structurally valid glyph may be misrecognized by an OCR system that is poorly suited to that style. OCR results should therefore be reported together with the recognition system and test conditions.

4.2.2. Skeleton and Topology Analysis

Skeleton-based methods represent a glyph using a simplified centerline structure, enabling the analysis of stroke connections, endpoints, intersections, and topological relationships. Such representations can help identify broken connections, redundant branches, and structural distortions that are difficult to describe through pixel-level similarity alone.

Research on skeleton extraction for hard-pen regular script illustrates the relevance of stroke characterization and ambiguous-zone detection to structural analysis [26]. Nevertheless, the reliability of skeleton-based evaluation depends on binarization, skeletonization, and the treatment of intersecting or overlapping strokes. These limitations are particularly important for calligraphic styles, in which stroke width, texture, and intentional connections may complicate the extraction of a stable skeleton.

Component-level and local-region assessment aims to detect errors that global image metrics can overlook, including missing or displaced radicals, abnormal stroke junctions, and local shape distortions. Some studies provide task-specific structure or component representations, but “stroke coverage,” “component matching,” and “component-position deviation” are not treated here as universally established metrics unless a cited study defines and validates them. In this review, these terms refer to potential indicators that future font benchmarks could operationalize with explicit formulas, reference annotations, and validation against expert judgments.

Component-level assessment is particularly relevant to scripts such as Chinese, where characters are composed of strokes and structural components. Existing font-generation studies have explored localized style representations and fine-grained spatial correspondence[27][28]. These approaches provide useful foundations for examining whether generated components preserve their appearance and spatial relationships.

Potential evaluation measures include stroke coverage, component matching, and component-position deviation. These terms describe possible ways of assessing structural completeness and spatial consistency; they should not be presented as established, standardized metrics unless a specific study formally defines and validates them. Their implementation would require explicit definitions of stroke or component correspondences, matching rules, and normalization procedures.

Local-region evaluation focuses on details such as stroke endpoints, turns, intersections, and component connections. The SSNet study uses structural modules and a dual-mask Hausdorff distance to constrain contour errors in Chinese typography generation[29]. This illustrates how localized structural information can be incorporated into generation or assessment. However, a method used as a generation constraint should not automatically be treated as an independently validated evaluation metric. The precise role of each method must be distinguished when comparing studies.

Optical character recognition (OCR) can support two related but distinct evaluation tasks. For character-content correctness, OCR is used to test whether a generated glyph is recognized as the intended character; the relevant outcome is character identity or recognition accuracy. For readability, OCR performance may be used as an indirect proxy for how reliably text can be read under specified size, contrast, or viewing conditions. OCR accuracy is not identical to human readability: it depends on the recognizer, training data, script, and test conditions, so it should be interpreted alongside human reading measures when the application requires them.

4.3. Style, Readability, and Aesthetic Assessment

Font quality extends beyond character correctness and structural integrity. A generated font must also preserve relevant stylistic characteristics, support its intended reading task, and meet the visual expectations of its users. These dimensions are related but should be evaluated separately because success in one does not guarantee success in the others.

4.3.1. Style Consistency

Style assessment examines whether generated glyphs preserve the intended design characteristics, such as stroke thickness, weight distribution, proportions, visual balance, curve characteristics, and decorative features. Depending on the task, evaluation may involve reference-based feature comparison, learned visual representations, semantic attributes, or human judgments.

Style consistency concerns whether generated glyphs match the intended typeface characteristics, including stroke weight, terminals, proportions, rhythm, and overall visual identity. FontCLIP provides a typography-specific vision-language representation that supports semantic style attributes and retrieval [4]. Its outputs can contribute evidence about style similarity, but it is not itself a comprehensive quality score and does not replace explicit structural or character-correctness checks.

4.3.2. Legibility and Readability

Legibility and readability are related but distinct concepts. Legibility concerns how clearly individual characters can be distinguished and recognized. Readability concerns the ease and efficiency with which readers can process text in a particular context, and may be influenced by font design, character spacing, text size, viewing conditions, and reading purpose.

Legibility refers to whether individual characters can be visually distinguished and identified; readability refers more broadly to how easily and efficiently text can be read in context. OCR recognition accuracy may help estimate content recognizability, while human reading speed, reading errors, critical print size, and preference can provide different evidence about readability and accessibility[30].

OCR accuracy can also be used as an automated proxy for character recognizability under specified conditions. However, it does not directly measure human reading speed, comprehension, comfort, or sustained reading experience. Therefore, OCR-based results should be described as machine-recognition performance unless a study has established their relationship with human readability judgments. Where practical readability is the target, OCR should be complemented by human evaluation.

4.3.3. Aesthetic Quality

Aesthetic assessment concerns subjective preferences and design qualities, including visual appeal, stylistic coherence, originality, and suitability for a particular application. Common approaches include expert ratings, user studies, Likert-scale ratings, mean opinion scores (MOS), and pairwise preference comparisons. Computational aesthetic models may provide additional predictions, but their relevance to font design depends on the training data, evaluation criteria, and relationship between model scores and human judgments.

Aesthetic preference is affected by cultural background, design experience, intended use, and individual taste. Consequently, a single aesthetic score should not be interpreted as a universally valid measure of font quality. Reports should identify the evaluator population, rating procedure, and criteria used. When comparing fonts intended for different purposes, application context should also be considered.

In summary, style consistency, legibility, readability, and aesthetics represent distinct but interacting aspects of font quality. A comprehensive assessment should state which of these dimensions is being measured and avoid using a proxy—such as OCR accuracy or visual-semantic similarity—as if it captured the entire user experience.

4.4. Multimodal and Large-Model Evaluation

Multimodal approaches combine visual information with text, semantic descriptions, or other contextual signals. They offer a potential means of assessing whether a generated font matches an intended style or design description, but their capabilities must be distinguished from those of dedicated font-quality metrics.

Vision-language models learn associations between visual representations and language. In typography-related applications, these representations can support font retrieval, attribute prediction, and comparisons between font images and textual descriptions. FontCLIP is a typography-specific example of this direction [4]. Such models can provide semantic evidence that complements pixel-level or structural analysis. Nevertheless, a high image-text similarity score does not by itself demonstrate that individual characters are correct, strokes are complete, or the font is readable.

The role of visual-text generation models also requires careful distinction. AnyText focuses on multilingual visual-text generation and editing rather than font-family generation [20]. It is relevant to evaluation of rendered text within images, but its task and success criteria differ from generating a reusable typeface or assessing a font family.

https://doi.org/10.67728/eci.2026.004

Large language models (LLMs) and vision-language models (VLMs) may support evaluation by integrating OCR outputs, structural measurements, and descriptions of intended styles. Such systems could generate explanations or flag likely errors, but these uses should be treated as emerging applications until their agreement with expert judgments, robustness, and sensitivity to typographic defects have been validated on suitable datasets.

The LLM-as-a-Judge paradigm has been studied in language and multimodal evaluation, but its applicability to font assessment requires task-specific validation [21][22]. Results may depend on prompts, model versions, image inputs, and scoring rubrics. A large model should therefore not be treated as a reliable judge merely because it can produce fluent explanations; its judgments need to be compared with human ratings and objective checks.

Accordingly, multimodal and large-model-based methods should currently be viewed as complementary approaches whose reliability depends on the task and evidence available. Their assessment should consider agreement with human judgments, sensitivity to structural defects, consistency across prompts or models, and performance across scripts and font styles. Claims about their effectiveness should be limited to the tasks and experimental settings actually evaluated in the cited studies.

4.5. Datasets and Evaluation Protocols

Datasets influence both the development of font-generation systems and the validity of quality assessment. Their purpose, composition, annotation scheme, and test protocol determine which conclusions can reasonably be drawn. The resources discussed below are distinguished by role: font/typeface collections, generation datasets, typography-attribute resources, general multimodal pretraining data, and dedicated quality-assessment datasets. These categories should not be treated as interchangeable.

Font-generation datasets provide glyph images, typeface samples, or character sets for model training and generation evaluation. Google Fonts provides a source of Latin typefaces, while Chinese-font studies may use open font collections such as Source Han Fonts or construct task-specific datasets. The availability of multiple typefaces and character sets supports the evaluation of style transfer and font generation, but the use of a generation dataset does not necessarily mean that it includes human quality annotations or a dedicated evaluation benchmark.

Font-attribute datasets provide labels or textual descriptions associated with font appearance. The FontCLIP Attribute Dataset, for example, supports typography-related semantic representation and attribute-oriented applications [4]. Such resources may facilitate style analysis and retrieval, but attribute labels should not be treated as direct measurements of character correctness, structural integrity, or readability unless those dimensions are explicitly annotated.

General AIGC image-quality datasets are designed for broad image-generation or image-text learning tasks rather than typography specifically. LAION, for example, is a large-scale image-text resource used in multimodal learning, not a dedicated AIGC image-quality benchmark and not a font-quality benchmark. Such resources may support representation learning, but they do not automatically provide annotations for character identity, glyph topology, readability, or typographic quality.

Table 3 summarizes representative resources and their roles. A resource is not described as a dedicated quality benchmark unless its task and annotations support that interpretation. Scale and annotation details are omitted where they cannot be established from the material reviewed; the table instead emphasizes each resource's primary role and its limitations for font-quality evaluation.

Table 3. Representative Datasets Relevant to Font Generation and Quality Assessment.

Resource	Dataset/Resource Role	Potential Use in This Review	Important limitation
Google Fonts	Typeface collection	Font generation and style-transfer experiments	Availability of typefaces does not imply quality annotations
Fonts Dataset (USC)	Synthetic font dataset	Font generation and representation-learning research	Evaluation suitability depends on the specific task and split
FontCLIP Attribute Dataset	Font-attribute and semantic resource	Style-attribute analysis, retrieval, and multimodal research	Attribute labels do not automatically measure overall font quality
Source Han Fonts	Typeface resource for Chinese scripts	Chinese-font generation and character rendering	A font collection is not, by itself, a quality benchmark
General AIGC image datasets (e.g., LAION)	General image-text learning	Contextual methods for image-text representation	Not designed specifically for glyph structure, character correctness, or readability

Source: Author.

These resources serve different purposes and should not be compared as though they were interchangeable benchmarks. Generation data, semantic-attribute annotations, and human quality judgments provide different evidence. A dedicated font-quality

benchmark would ideally specify script and language coverage, typeface diversity, glyph-level labels, structural defect annotations, human ratings where relevant, train/test separation, and a reproducible evaluation protocol.

Across the four categories, a consistent pattern emerges: image-level metrics offer efficient baselines; character- and structure-aware methods address typography-specific constraints; human-centred methods capture readability and aesthetic preference; and multimodal or large-model approaches may connect visual evidence with semantic descriptions. No category is sufficient by itself. The following tables map representative studies and metrics to the analytical framework. The synthesis is qualitative and does not imply that every method has been validated on a common benchmark.

5. Discussion

5.1. Current Challenges

Within the literature reviewed here, font-quality assessment appears less standardized than font-generation research. Three recurring challenges are evident in the studies discussed below.

First, dedicated typography-specific metrics are lacking. Many studies still use PSNR, SSIM, FID, and KID, which primarily capture pixel differences or global feature distributions. These metrics cannot adequately describe the hierarchical strokes, radical structures, and semantic constraints characteristic of fonts. When generated fonts contain local missing strokes, component displacement, or structural deformation, conventional metrics may fail to reflect practical quality, producing results that diverge from human perception. Although FontCLIP introduces semantic font representations, its main applications remain font retrieval and style analysis rather than standardized automatic quality assessment.

Second, the reviewed studies use datasets and protocols that vary in font types, character coverage, annotation schemes, and evaluation procedures. This variation limits direct comparison and indicates a need for more consistently reported benchmarks. The present review does not establish that no unified resource exists anywhere; rather, a single widely adopted benchmark covering all relevant quality dimensions was not identified in the sources examined.

Third, the reliability of large-model evaluation remains uncertain. LLM-as-a-Judge has demonstrated scalability in text, image, and multimodal evaluation, but studies have identified prompt sensitivity, model bias, limited consistency, and restricted cross-task generalization[31]. Different models, or even the same model under different prompts, may produce substantially different judgments, affecting stability and credibility. Font assessment is even more difficult because a model must understand not only character content but also typography structure, style, weight, and layout. Building reliable large-model evaluation systems is therefore a central challenge.

5.2. Future Research Directions

The first priority is to establish a unified multidimensional benchmark for font quality. Future systems should move beyond conventional image-quality metrics and integrate character correctness, stroke completeness, skeleton topology, component relationships, style consistency, readability, and aesthetic quality. A public benchmark should specify standardized dataset splits, evaluation protocols, and reporting requirements so that algorithms can be compared fairly. Drawing on the broader development of AI benchmarks, future datasets should also cover multiple languages, typefaces, and application scenarios to improve generalization and international applicability.

The second priority is to improve the reliability and interpretability of large-model evaluation. Large models can jointly analyze character structure, font style, and textual semantics and can generate interpretable reports and design suggestions. However, evaluation consistency, model bias, prompt robustness, and cross-model stability must be systematically addressed. Recent LLM-as-a-Judge research suggests combining dedicated reliability benchmarks with human evaluation to establish human-machine collaborative assessment [32][33]. For typography, domain knowledge should be incorporated explicitly, including stroke hierarchy, component relationships, and writing rules. This would enable an evaluation framework that closes the loop among generation, assessment, explanation, and feedback-based optimization.

5.3. Implications for Research and Application

The reviewed evidence suggests that font evaluation should be treated not merely as a final scoring stage but as an active component of the generation pipeline. Objective structural metrics can provide stable optimization signals, semantic models can evaluate style and content consistency, and human feedback can calibrate aesthetic judgments. A layered framework could therefore support model training, model selection, error diagnosis, and downstream font recommendation. Such a framework would also make evaluation more useful for real design workflows, where the key question is not only whether a generated glyph is similar to a reference, but whether it is structurally valid, stylistically coherent, readable, aesthetically appropriate, and suitable for a particular cultural or commercial context.

https://doi.org/10.67728/eci.2026.004

A further issue concerns the relationship between objective metrics and human judgments. Objective metrics offer scalability and reproducibility, but they can encode assumptions that do not correspond to typographic perception. Human evaluation captures contextual and aesthetic preferences, yet it is expensive and can vary across evaluators[34]. A more robust design is therefore a hierarchical evaluation pipeline in which automatic metrics first detect basic content and structural errors, semantic models assess style and meaning, and human judgments calibrate higher-level aesthetic and application-specific criteria. Such a pipeline can also generate richer supervision for future learning-based evaluators.

Cross-lingual and cross-cultural evaluation should receive particular attention. A metric trained mainly on Latin fonts may not capture the structural logic of Chinese characters, Japanese scripts, or other writing systems. Likewise, judgments of elegance, formality, visual balance, and decorative appropriateness can depend on cultural and application contexts. Multilingual datasets and culturally diverse human annotations are therefore necessary if future evaluation systems are expected to operate across international typography applications.

Finally, benchmark design should reflect realistic generation scenarios rather than only isolated glyph-level comparison. Test suites can include few-shot generation, unseen-character generation, style transfer, multilingual generation, and text-image editing, together with controlled structural defects and human preference judgments. Standardized error taxonomies would make it possible to report not only a single aggregate score but also the types and locations of failures. This would make evaluation more actionable for model developers and designers.

The practical value of a unified benchmark also lies in its ability to support error diagnosis. Instead of treating quality as a single scalar, evaluation should report interpretable sub-scores and localized failure information. For example, a system may achieve high visual similarity while receiving low structural-integrity scores because of a missing radical, or it may preserve structure while receiving a lower style-consistency score because the target weight or stroke rhythm has drifted. Such multidimensional reporting would allow developers to identify whether an observed weakness originates from content preservation, structural modeling, style transfer, readability, or aesthetic control. It would also make the evaluation results easier for designers and end users to understand.

Illustrative example (not an experiment reported in this review): if a generated Chinese glyph receives a high SSIM score but its OCR output differs from the target character, the result should be flagged as a content-correctness failure despite visual similarity. If OCR matches the target but skeleton analysis detects a broken stroke or missing component, the result should be flagged as a structural failure. If content and structure pass but style similarity is low, a typography-specific semantic model and a blinded human rating can help determine whether the discrepancy reflects intended stylistic variation or undesirable style drift. This example shows why the proposed pipeline reports separate evidence and sub-scores rather than a single unqualified quality score.

Another promising direction is to use evaluation results as feedback for generation. Once structural and semantic errors can be detected automatically, evaluation models can provide targeted signals for model fine-tuning or iterative correction. Human preference data can further calibrate these signals for aesthetic and application-specific judgments. In this sense, the ultimate goal is not merely to rank generated fonts but to establish a feedback mechanism through which generation systems learn from evaluation and progressively improve. This generation–evaluation–optimization loop may become a core component of future intelligent font-design platforms.

5.4. Font Quality Annotation Dimensions

A scientific and unified font-quality evaluation system requires multidimensional annotation. Current studies commonly consider character correctness, structural integrity, readability, stylistic consistency, and aesthetic quality. Character correctness is foundational and focuses on whether a generated glyph expresses the intended semantics; OCR accuracy and manual verification can be combined, with additional checks for missing or incorrect strokes and character confusions in Chinese-font generation. Stroke and structural integrity concern geometric correctness, including stroke connections, skeleton topology, component positions, and overall layout. Readability evaluates recognition efficiency in realistic reading scenarios using OCR accuracy, human recognition accuracy, and reading time[36]. Human experiments reflect actual user experience, whereas OCR supports large-scale testing. Style consistency concerns preservation of target design characteristics such as stroke thickness, weight, center of gravity, and decorative style; multimodal attribute labels can support a transition from visual similarity toward semantic consistency. Aesthetic quality concerns artistic expression and user preferences and is commonly assessed through expert ratings and user-based MOS, while also requiring consistency of design language among different characters in the same font family[36].

5.5. Evaluation Experiments and Reporting Protocols

For reproducibility and comparability across studies, font-quality assessment should use standardized experimental and reporting protocols. Existing work generally combines objective and subjective evaluation. Objective measures include PSNR, SSIM, FID, KID, and OCR accuracy, enabling rapid assessment of model performance. Subjective assessment uses expert ratings or user

experiments to evaluate style, attractiveness, and readability. These approaches complement each other and provide a more complete representation of actual font quality.

Experimental reports should consistently specify dataset sources, training-set size, test-set partitioning, evaluation metrics, and experimental environments. Results should separately report character correctness, structural integrity, style consistency, and subjective scores and should use representative examples to explain model strengths and weaknesses. For multimodal evaluation, reports should additionally specify the evaluation model, prompt design, and evaluation procedure to improve transparency and reproducibility. In the future, evaluation protocols are expected to evolve from single-metric assessment toward multidimensional frameworks integrating objective metrics, structural analysis, semantic understanding, and human perception[5].

6. Conclusion

Following the research line from generation to evaluation, this review examined in this structured narrative review the development, key technologies, and future trends of quality assessment for AI-generated fonts. At the generation level, font generation has progressed from early GAN-based end-to-end style transfer toward a diverse landscape combining structure-aware modeling, diffusion-based generation, and semantically driven multimodal large models. The structural complexity and stylistic diversity of generated fonts have consequently increased. In parallel, quality assessment has expanded from conventional image metrics such as PSNR, SSIM, and FID to multidimensional evaluation covering character correctness, skeleton topology, components, style consistency, readability, and aesthetics. FontCLIP and multimodal large models have further introduced semantic and intelligent evaluation capabilities, indicating a broader transition from pixel-level comparison toward semantic and human-centered assessment.

The evidence synthesized in this review indicates that no single metric among those discussed captures all relevant aspects of AI-generated font quality. Pixel-level metrics may miss character or component errors; distribution-level metrics do not directly assess individual character identity; and human evaluation, while important for perception and aesthetics, can be costly and variable. A practical assessment protocol should therefore combine complementary evidence, report its limits, and validate proposed indicators against typography-specific data and human judgments.

Overall, the field is moving from a fragmented collection of image metrics toward an integrated evaluation science for typography. Achieving this transition will require cooperation among computer vision, natural-language processing, typography, design, and human-computer interaction research. The resulting evaluation infrastructure can serve not only academic benchmarking but also practical quality assurance and intelligent design, helping ensure that increasingly powerful font-generation systems remain accurate, controllable, interpretable, and aligned with human needs.

Funding: This research received no external funding. The APC was waived by the journal.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new datasets were generated or analyzed during this study.

Conflicts of Interest: The authors declare no conflicts of interest.

References

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, et al., “Generative adversarial networks,” *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020, doi: 10.1145/3422622.

[2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, et al., “Learning transferable visual models from natural language supervision,” in *Proc. 38th Int. Conf. Mach*. Learn. *(ICML)*, 2021, pp. 8748–8763.

[3] R. Rombach, B. Blattmann, D. Lorenz, P. Esser, and B. Ommer., “High-resolution image synthesis with latent diffusion models,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 10684–10695, doi: 10.1109/cvpr52688.2022.01042.

[4] Y. Tatsukawa, I.-C. Shen, A. Qi, Y. Koyama, T. Igarashi, and A. Shamir, “FontCLIP: A semantic typography visual-language model for multilingual font applications,” *Comput. Graph. Forum*, vol. 43, no. 2, Art. no. e15043, 2024, doi: 10.1111/cgf.15043.

[5] L. Wang, Y. Liu, M. Y. Sharum, R. Yaakob, K. A. Kasmiran, and C. Wang, “Deep learning for Chinese font generation: A survey,” *Expert Syst. Appl.*, vol. 276, Art. no. 127105, 2025, doi: 10.1016/j.eswa.2025.127105.

[6] M. Heusel, H. Ramsauer, H. Unterthiner, B. Nessler, and S. Hochreiter, “GANs trained by a two time-scale update rule converge to a local Nash equilibrium,” in *Adv. Neural Inf. Process. Syst*. 30, 2017, pp. 6626–6637.

[7] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004, doi: 10.1109/TIP.2003.819861.

[8] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in Proc. *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 586–595, doi: 10.1109/CVPR.2018.00068.

[9] J. Yuan, X. Cao, L. Cao, J. Lin, and X. Cao, "PSCR: Patches sampling-based contrastive regression for AIGC image quality assessment," *arXiv preprint arXiv:2312.05897*, 2023, doi: 10.48550/arXiv.2312.05897.

[10] J. Yuan, X. Cao, J. Che, Q. Wang, S. Liang, W. Ren, et al., "TIER: Text-image encoder-based regression for AIGC image quality assessment," *arXiv preprint arXiv:2401.03854*, 2024, doi: 10.48550/arXiv.2401.03854.

[11] S. Azadi, M. Fisher, V. G. Kim, Z. Wang, E. Shechtman, and T. Darrell, "Multi-content GAN for few-shot font style transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 7564–7573, doi: 10.1109/CVPR.2018.00789.

[12] S. Park, S. Chun, J. Cha, B. Lee, and H. Shim, "Multiple heads are better than one: Few-shot font generation with multiple localized experts," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 13900–13909. doi: 10.1109/iccv48922.2021.01364.

[13] Y. Xie, X. Chen, L. Sun, and Y. Lu, "DG-Font: Deformable generative networks for unsupervised font generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 5126-5136, doi: 10.1109/CVPR46437.2021.00509.

[14] J. Cha, S. Chun, G. Lee, B. Lee, S. Kim, and H. Lee, "Few-shot compositional font generation with dual memory," in *Computer Vision – ECCV 2020*, ser. Lecture Notes in Computer Science, vol. 12364, A. Vedaldi, H. Bischof, H. Brox, and J. M. Frahm, Eds. Cham, Switzerland: Springer, 2020, doi: 10.1007/978-3-030-58529-7_43.

[15] Y. Gao, Y. Guo, Z. Lian, Y. Tang, and J. Xiao, "Artistic glyph image synthesis via one-stage few-shot learning," *ACM Trans. Graph.*, vol. 38, no. 6, Art. no. 185, 2019, doi: 10.1145/3355089.3356574.

[16] J. Park, A. U. Hassan, and J. Choi, "CCFont: Component-based Chinese font generation model using generative adversarial networks (GANs)," *Appl. Sci.*, vol. 12, no. 16, Art. no. 8005, 2022, doi: 10.3390/app12168005.

[17] X. Chen, L. Wu, Y. Su, L. Meng, and X. Meng, "Font transformer for few-shot font generation," *Comput. Vis. Image Underst.*, vol. 245, Art. no. 104043, 2024, doi: 10.1016/j.cviu.2024.104043.

[18] A. Fu, F. Yu, A. Liu, Z. Wang, J. Wen, J. He, and Y. Qiao, "Generate like experts: Multi-stage font generation by incorporating font transfer process into diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 6892–6901, doi: 10.1109/CVPR52733.2024.00658.

[19] Z. Yang, D. Peng, Y. Kong, Y. Zhang, C. Yao, and L. Jin, "FontDiffuser: One-shot font generation via denoising diffusion with multi-scale content aggregation and style contrastive learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 7, pp. 6603–6611, 2024, doi: 10.1609/aaai.v38i7.28482.

[20] Y. Tuo, W. Xiang, J. Y. He, Y. Geng, and X. Xie, "AnyText: Multilingual visual text generation and editing," in *Proc. 12th Int. Conf. Learn. Represent. (ICLR)*, 2024, pp. 56783–56799.

[21] J. Gu, X. Jiang, Z. Shi, H. Tan, X. Zhai, C. Xu, et al., "A survey on LLM-as-a-Judge," *The Innovation*, vol. 7, no. 6, Art. no. 101253, 2026, doi: 10.1016/j.xinn.2025.101253.

[22] A. Li, B. Jiang, L. Huang, A. Beigi, C. Zhao, Z. Tan, et al., "From generation to judgment: Opportunities and challenges of LLM-as-a-Judge," in *Proc. 2025 Conf. Empirical Methods Nat. Lang. Process. (EMNLP)*, 2025, pp. 2757–2791, doi: 10.18653/v1/2025.emnlp-main.138.

[23] H. Zhou, A. Yu, L. Kang, Y. Fan, Y. Fan, X. Song, et al., "Reading ≠ seeing: Diagnosing and closing the Typography Gap in vision-language models," arXiv preprint arXiv:2603.08497, 2026, doi: 10.48550/arXiv.2603.08497.

[24] A. Hore and D. Ziou, "Image quality metrics: PSNR vs. SSIM," in *Proc. 20th Int. Conf. Pattern Recognit. (ICPR)*, 2010, pp. 2366–2369, doi: 10.1109/ICPR.2010.579.

[25] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying MMD GANs," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018, arXiv preprint arXiv:1801.01401v4.

[26] Z. Xu, F. Qin, N. Xiong, H. Shi, J. Zhang, and W. Lin, "Skeleton extraction of hard-pen regular script based on stroke characterization and ambiguous zone detection," *IET Image Process*., vol. 18, no. 6, pp. 1504–1515, 2024, doi: 10.1049/ipr2.13044.

[27] S. Park, S. Chun, J. Cha, B. Lee, and H. Shim, "Few-shot font generation with localized style representations and factorization," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 3, pp. 2393–2402, 2021, doi: 10.1609/aaai.v35i3.16340.

[28] L. Tang, Y. Cai, J. Liu, Z. Hong, M. Gong, M. Fan, et al., "Few-shot font generation by learning fine-grained local styles," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 7885-7894, doi: 10.1109/CVPR52688.2022.00774.

[29] X. Yang, S. Wang, Y. Zhang, et al., "SSNet: Structure-semantic net for Chinese typography generation based on image translation," *Neurocomputing*, vol. 371, pp. 15–26, 2020, doi: 10.1016/j.neucom.2019.08.072.

[30] M. Ohnishi and K. Oda, "The effect of style and weight on the readability of Japanese fonts: A study of reading evaluations with standard off-the-shelf fonts," *J. Illuminating Eng. Inst. Japan*, vol. 101, no. 10, pp. 474–483, 2017, doi: 10.2150/jieij.101.474.

[31] T. R. McIntosh, T. Susnjak, N. Arachchilage, T. Liu, P. Watters, and M. N. Halgamuge, "Inadequacies of large language model benchmarks in the era of generative artificial intelligence," *IEEE Trans. Artif. Intell.*, vol. 7, no. 1, pp. 22–39, 2026, doi: 10.1109/TAI.2025.3569516.

[32] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, et al., "A survey on evaluation of large language models," *ACM Trans. Intell. Syst. Technol.*, vol. 15, no. 3, Art. no. 39, pp. 1–45, 2024, doi: 10.1145/3641289.

[33] S. Minaee, T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain, et al., "Large language models: A survey," arXiv preprint arXiv:2402.06196, 2024.

[34] M. T. R. Laskar, S. Alqahtani, M. S. Bari, M. Rahman, M. A. M. Khan, H. Khan, et al., "A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations," in *Proc. 2024 Conf. Empirical Methods Nat. Lang. Process. (EMNLP)*, Miami, FL, USA, 2024, pp. 13785–13816, doi: 10.18653/v1/2024.emnlp-main.764.

[35] G. E. Legge, Y. Xiong, Q. Gao, R. Gage, T. Knickel, and C. Bigelow, "Assessment of newly designed fonts for visual accessibility," *PLoS One*, vol. 21, no. 3, Art. no. e0345068, Mar. 2026, doi: 10.1371/journal.pone.0345068.

[36] Y. Deng, C. C. Loy, and X. Tang, "Image aesthetic assessment: An experimental survey," *IEEE Signal Process. Mag.*, vol. 34, no. 4, pp. 80–106, 2017, doi: 10.1109/MSP.2017.2696576.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of Apollo and/or the editor(s). Apollo and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.