# Characterizing Visual Narrative Freedom under Loose Image–Text Alignment

**Yanru Jiang**[1] **Gavin Olson**[1] **Eugenio Herrera-Berg**[2]
**Rick Dale**[1] **Hongjing Lu**[1*] **Elisa Kreiss**[1*]
[1]University of California, Los Angeles [2]CENIA
{yanrujiang, gavinolson, rdale, hongjing, ekreiss}@ucla.edu
eugenio.herrera@cenia.cl [*]Corresponding author(s)

## Abstract

Mapping linguistic to visual representations is a central objective in vision-language model (VLM) development. Consequently, current VLM benchmarks and training objectives predominantly optimize for literal, descriptive correspondence between modalities. However, in real-world multimedia contexts, text and images are rarely strictly redundant; they are often loosely coupled, interacting at higher narrative levels. To capture this dynamic, we introduce *Visual Narrative Freedom* (VNF): the degree to which a generative system can produce multiple plausible visual realizations from an underdetermined textual input. In this paper, we systematically evaluate text-to-image (T2I) generation across a continuum of linguistic constraints, conditioning models on visual descriptions, captions, and full news articles. Using perceptual, structural, and semantic metrics, we first demonstrate that as textual constraints loosen (and VNF increases), T2I models produce images that are significantly more diverse from ground-truth reference photographs. Human evaluation further reveals that providing generation models with greater visual narrative freedom significantly increases the likelihood that their outputs are preferred over authentic news photographs, despite the known stylistic limitations of AI imagery. Ultimately, these findings suggest that the creation of persuasive multimodal misinformation is more imminent than evaluations of T2I systems based solely on text descriptions may indicate. This highlights the need for VLM evaluation frameworks that better capture the underdetermined nature of real-world image-text alignment.

## 1 Introduction

Mapping linguistic representations to visual representations and integrating the two modalities has been at the forefront of recent vision–language model (VLM) development (Zong et al., 2023; Li et al., 2021). Many existing approaches align images and text through shared embeddings or cross-modal supervision (Li and Tang, 2026; Fei et al., 2022), often optimizing for literal correspondence between textual descriptions and visual content (Radford et al., 2021). However, such formulations overlook a fundamental aspect of multimodal communication: images and text do not merely translate information but interact at higher narrative and symbolic levels (Zwaan, 2014; Cheema et al., 2023; Barthes, 1977).

While image-to-text (ITT) and text-to-image (TTI) systems are typically studied separately due to differences in architectures and training objectives, both aim to model relationships between visual and linguistic modalities (Zhang et al., 2024). In this work, we use the term VLMs broadly to refer to models that connect visual and textual modalities, including image captioning, image generation, and cross-modal retrieval. This grouping shifts the focus from architectural distinctions to **how image–text relationships are structured and operationalized in current VLMs**, making our analysis relevant across the broader spectrum of vision–language applications.

In many multimedia contexts, images do not function as literal visual translations of text (Zwaan, 2014; Lommatzsch et al., 2022). Across domains such as journalism, advertising, and social media, images often complement or symbolically represent textual narratives rather than redundantly mirroring them (Cheema et al., 2023). As a result, the relationship between a text and an image may be loosely coupled: a single textual narrative can plausibly correspond to multiple visual realizations. Generative systems can exploit this flexibility because, unlike real-world photography, they are not constrained by physical access, timing, or the availability of real-world scenes (Ciammaichella, 2023).

We refer to this generative property as *Visual Narrative Freedom* (VNF): the degree to which a system can produce multiple plausible visual real-

izations for the same textual context. VNF emerges when the mapping from text to image is underdetermined, allowing generative models to instantiate different visualizations of the same narrative. In practice, loosely coupled inputs such as full articles afford greater VNF than tightly specified descriptions.

Visual journalism provides a particularly high-stakes context in which these dynamics become visible (Strikovic and Cools, 2025; Paik et al., 2023). Images accompanying news articles often serve narrative or framing functions rather than strictly documenting the events, locations, or figures described in the text. At the same time, news photographs carry evidentiary authority and are often perceived as eyewitness records of reality (Banks, 2013; Ristovska, 2020; Rubinstein and Sluis, 2008). These characteristics make news media an informative setting for studying the downstream impact of generative systems when image–text relationships are loosely constrained. Even when generated images achieve high perceptual fidelity, they may retain stylistic cues that distinguish them from traditional photojournalism (Strikovic and Cools, 2025; Thomson et al., 2025), potentially resulting in an inherent dispreference in journalistic contexts.

News articles serve as an especially useful resource to investigate the effects of VNF, since they inherently provide naturalistic examples of text that vary across three VNF levels for each individual image: *descriptions* (which can be generated from the image directly when absent), *captions*, and full news *articles*. This allows us to examine how different levels of textual constraint influence the diversity of generated images and how these differences interact with audience preferences for VNF alignment when images are presented alongside their corresponding articles.

In summary, our main contributions are: (1) We show that loosely coupled image–text relationships is not solely an image-to-text (I2T) problem but generalizes to evaluating text-to-image (T2I) systems, making it a bidirectional Vision-Language alignment issue. (2) We introduce VNF, a measurable generative property that characterizes the range of plausible visual realizations from a given textual context. (3) Through pre-registered experiments in news media, we demonstrate that VNF increases audience preference for images presented alongside articles, even when AI-generated imagery carries stylistic disadvantages.

## 2 Related Work

Our work connects two strands of literature on image–text relationships. The first examines how VLMs operationalize relationships between images and text, typically assuming tightly aligned descriptive correspondence. The second comes from multimedia communication and journalism, where images and text often interact more loosely in real-world settings. Building on these perspectives, we introduce VNF as a property that emerges when generative models operate under loosely constrained textual inputs.

### 2.1 Image–Text Relationships in Vision–Language Models

Across VLM tasks, the dominant paradigm assumes relatively tight semantic alignment between images and text (Fei et al., 2022). Image captioning datasets typically describe visible entities, attributes, and actions within an image (Kreiss et al., 2022a), and many TTI benchmarks similarly rely on prompts that explicitly specify the visual scene to be generated (Baraheem et al., 2023). This assumption is reflected in both training objectives and dataset construction. Large-scale contrastive frameworks such as CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021), CLOOB (Shin et al., 2022), and DeCLIP (Li et al., 2022) learn multimodal representations by maximizing correspondence between paired images and captions. Widely used benchmarks, including MSCOCO (Lin et al., 2014), Flickr30K (Young et al., 2014), Visual Genome (Krishna et al., 2017), and Conceptual Captions (Sharma et al., 2018), similarly pair images with text that directly describes visible content.

Recent work has begun to explore richer forms of cross-modal interaction. For example, the Visual Commonsense Reasoning dataset (Zellers et al., 2018) evaluates models' ability to infer motivations and social context from images, while Concadia (Kreiss et al., 2022b) distinguishes between *descriptions*, which substitute for images by focusing on visible features, and *captions*, which accompany images to provide contextual information. Similarly, BriVL (Huo et al., 2021) introduces a weakly correlated vision–language pretraining framework, relaxing the assumption that paired image–text data must exhibit strict semantic correspondence.

At the same time, modern multimodal architectures increasingly support extended textual conditioning beyond short captions (Liu et al., 2024;

Zhang et al., 2024). However, evaluation frameworks still emphasize literal correspondence between textual descriptions and visual content (Lommatzsch et al., 2022). This paradigm underrepresents the broader range of relationships found in real-world multimedia communication, where images and text may interact through complementary, loosely aligned, or even contrasting narrative roles.

### 2.2 Image–Text Relationships in Multimedia Communication

In many real-world multimedia contexts, images and text interact through relationships that extend beyond literal description. For example, social media posts frequently pair images with captions that frame interpretation rather than strictly describe visual content (Mehmet and Clarke, 2016). In news media, images accompanying an article may illustrate broader themes, frame particular actors, or evoke emotional responses without depicting the specific events described in the text (Lommatzsch et al., 2022; Oostdijk et al., 2020). In such contexts, audiences do not simply evaluate whether images correspond with text; instead, they interpret how the two modalities interact by drawing on contextual knowledge and narrative cues.

In news image–article pairings, this loosely coupled relationship is often described as the *depiction gap*: the difference between what an image literally depicts and why it is connected to the accompanying text (Lommatzsch et al., 2022). Depiction gaps arise because news images may be selected for thematic relevance, narrative framing, or practical constraints such as image availability and newsroom sourcing limitations (Cheema et al., 2023; Oostdijk et al., 2020). When such gaps are substantial, multiple distinct visual realizations may plausibly accompany the same narrative context, creating interpretive latitude in how textual narratives are visually represented. This flexibility motivates our investigation of VNF in TTI generation.

## 3 Characterizing Visual Narrative Freedom

To characterize VNF in TTI systems, we analyze generated images under three prompting conditions that share the same underlying context: *description*, *caption*, and *article*. These prompts vary systematically in the degree of semantic constraint imposed by the textual input.

### 3.1 VNF Latitude across TTI Tasks

These three common TTI prompting forms vary in visual specification, forming a continuum of textual constraint—from tightly specified descriptions to broader narrative context—and enabling us to examine how prompt scope shapes VNF in generative systems.

**Descriptions** represent the most restrictive condition, explicitly specifying visual attributes such as subjects, environment, and scene composition, leaving relatively little ambiguity about the intended image.

**Captions** provide moderate constraint, summarizing the depicted event while omitting many visual details and allowing for multiple plausible realizations.

**Articles** introduce the least restrictive condition and represent the novel condition introduced in our study. This condition is enabled by recent advances in TTI generators capable of processing longer textual inputs. Full articles provide rich narrative context but rarely specify the precise visual scene accompanying the story, affording greater interpretive latitude for image generation.

### 3.2 Measuring VNF

Because VNF concerns the range of visual interpretations arising from the same textual context, we measure variation directly in generated images. Using computational similarity metrics, we quantify this variation by comparing plausible generated image candidates from the same textual input to one another (*Intra-Prompt Diversity*) or to a reference photograph representing the ground-truth image pairing instance (*Reference Deviation*).

Across these measurements, if the VNF property holds, the diversity of generated images across candidates and their visual deviation from the original image should increase progressively from *description*- to *caption*- to *article*-based generation.

**Intra-Prompt Diversity** measures the diversity of visual realizations produced from the same prompt. For each textual input, multiple candidate images are generated and compared using pairwise similarity. Higher scores indicates greater diversity in visual interpretations and thus greater VNF. Formally, diversity is computed as the average pairwise distance (i.e., $1 - \mathrm{sim}$) across generated candidates:

$$\frac{1}{N(N-1)} \sum_{i \neq j} (1 - sim(I_i, I_j))$$

where $N$ is the number of generated images for a given prompt, and $I_i, I_j$ denote generated images from the same prompt.

**Reference Deviation** measures how far generated images deviate from the original photograph paired with each news instance, which serves as one possible visual realization of the narrative. Formally, we compute deviation as $1 - \text{sim}(I_{generated}, I_{reference})$, where higher scores indicate greater divergence from the original depiction and thus greater VNF.

**Similarity Metrics** To capture visual similarity across different levels of representation, from low-level perceptual features to high-level semantic concepts, we consider four complementary similarity metrics across perceptual, structure, semantics and conceptual:

- **LPIPS (Learned Perceptual Image Patch Similarity)** computes distances between intermediate feature activations from a pretrained AlexNet network, calibrated to correlate with human perceptual similarity judgments (Zhang et al., 2018).
- **DISTS (Deep Image Structure and Texture Similarity)** compares structural correlations and texture statistics derived from feature maps of a pretrained VGG network, capturing both spatial layout and texture similarity (Ding et al., 2022).
- **VGG Feature Similarity** measures similarity between deep visual representations extracted from a pretrained VGG network. Feature maps are globally pooled via adaptive average pooling and compared using cosine similarity to produce image-level embeddings capturing high-level visual structure (Simonyan and Zisserman, 2015).
- **CLIP Embedding Similarity** measures semantic similarity using cosine distances between image embeddings produced by the LongCLIP `ViT-L/14` model in a joint vision–language representation space aligned with textual concepts (Radford et al., 2021).

## 4 Model Experiments

To characterize the latitude of VNF in real-world applications, we construct a balanced set of 108 news instances drawn from five major outlets for both model experiment and human evaluation in a naturalistic context. The dataset is intentionally kept small, as the same stimuli are reused for human experiments and each instance is generated under multiple conditions, resulting in 324 unique text-image pairs.

### 4.1 Dataset

To assess the effect of VNF on TTI systems, we curated a dataset from two established multimodal news corpora: VisualNews (Liu et al., 2021), including four major outlets (*BBC*, *The Guardian*, *USA Today*, and *The Washington Post*), and *VOA News* from M2E2 (Li et al., 2020). Each selected item contained a news image, caption, and full article. To ensure experimental consistency, account for LongCLIP context window limitations (i.e., 248 tokens), and control participant engagement time, we restricted articles to 100–200 words, required captions to contain at least five words to avoid underspecified prompts, and conducted sampling to roughly reflect the underlying distribution of outlets and major news topics. After filtering and balancing, the final stimulus set contained 108 news instances. Topic distributions and token counts are reported in Appendix A.

**Generating Image Descriptions** For each instance, we derived an additional tightly constrained condition by generating a highly detailed image description using `GPT-5` (note that `GPT-4o` was used for the human experiment in Section 5.1, as `GPT-5` had not yet been released at that time). Unlike standard short alt-text descriptions commonly found online, these descriptions were deliberately constructed to be maximally specific, explicitly specifying subjects, environment, lighting, mood, and camera perspective. This condition creates the most constrained setting. The exact prompt is provided in the Appendix B.

### 4.2 Experiment Setup

For each news instance, we generated images under three conditions using prompts derived from the same underlying content: a description-based condition, a caption-based condition, and an article-based condition.

The generated image aspect ratio was selected to approximate the original photograph whenever possible (e.g., 16:9 or 4:3 $\rightarrow$ 1,536×1,024; 9:16 or 3:4 $\rightarrow$ 1,024×1,536), based on the closest available aspect-ratio options provided by the TTI systems. Each prompt generates 5 candidates per model, yielding an expected total of $108\ instances \times 3\ conditions \times 5\ candidates \times 3\ models = 4{,}860\ images$. In total, we generated 4,771 images, due to `Nano-Banana-2` refusals and the open-source model occasionally generating news article layouts rather than images under the article-based condition.

### 4.3 Model Selection

For this analysis, we prioritized generation speed and high-level compositional and semantic variation over perceptual fidelity. We considered three models: `Nano-Banana-2`, `Realistic-Vision-V3.0`, and `Qwen-Image`. `Nano-Banana-2` is a proprietary T2I model from the Imagen family (used for high-fidelity generation in human evaluations). `Realistic-Vision-V3.0` is a Stable Diffusion–based model fine-tuned for photorealistic generation; we used 50 inference steps and replace its CLIP backbone with LongCLIP (`ViT-L/14`) to support longer textual inputs (up to 248 tokens). `Qwen-Image` is an open-weight T2I model; we used the `DFloat11/Qwen-Image-DF11` checkpoint, a quantized variant optimized for memory-efficient deployment via CPU offloading, with 30 inference steps and a `true_cfg_scale` of 3.0. All other configurations used model defaults.

### 4.4 Results

Similarity metrics (LPIPS, DISTS, VGG, and CLIP) introduced in Section 3.2 are used for both intra-prompt diversity and reference deviation measurements, capturing variation across perceptual, structural, semantic, and conceptual dimensions.

**Intra-Prompt Diversity** According to Figure 1, at the aggregated level (i.e., diversity computed across image candidates for each prompt across all three models), the three generation conditions exhibit a clear trend: similarity decreases from description to caption to article, indicating increasing generation diversity under less constrained prompts. When examining individual models, this pattern is most consistent in `Qwen-Image`, which shows a strong monotonic decrease across all metrics. `Nano-Banana-2` follows a similar but more moderate trend. In contrast, `Realistic-Vision-V3.0` exhibits a partial deviation: caption-based generations show slightly higher similarity than description-based generations, indicating lower diversity under caption prompts. Nevertheless, its article-based generation still produces the lowest similarity (i.e., highest diversity). These patterns are consistently observed across all similarity measures.

**Reference Deviation** According to Figure 2, at the aggregated level (i.e., deviation computed across image candidates for each prompt across all three models), the three generation conditions exhibit a clear trend: the average similarity between generated images and reference images decreases from description- to caption- to article-based generation. Examining individual models, this pattern is most consistent in `Qwen-Image` and `Nano-Banana-2`, both of which show a strong monotonic decrease across all metrics. `Realistic-Vision-V3.0`, however, exhibits a slightly inconsistent pattern across metrics, with only DISTS and VGG showing a monotonic decrease. Nevertheless, similar to the intra-prompt diversity results, its article-based generations show the largest reference deviation. These patterns are largely consistent across similarity measures.

**Overall Results** Across all representation levels, we observe a consistent increase in both intra-prompt diversity and reference deviation as the generation prompt expands from description to caption to article. This pattern suggests that increasingly loose image–text coupling affords greater VNF within the same TTI model. This VNF property emerges across the full spectrum of visual representations, spanning perceptual, structural, semantic, and conceptual dimensions, thereby enabling the generator to produce progressively more diverse visual interpretations of the underlying event.

## 5 Human Evaluation

We recruit human participants to assess preferences for images generated under the three conditions, each compared against the journalist-selected reference. These evaluations reveal the audience-end implications of modern TTI systems for loosely coupled image–text tasks: as models achieve high visual fidelity in photorealistic images, they can effectively leverage VNF, shaping how audiences

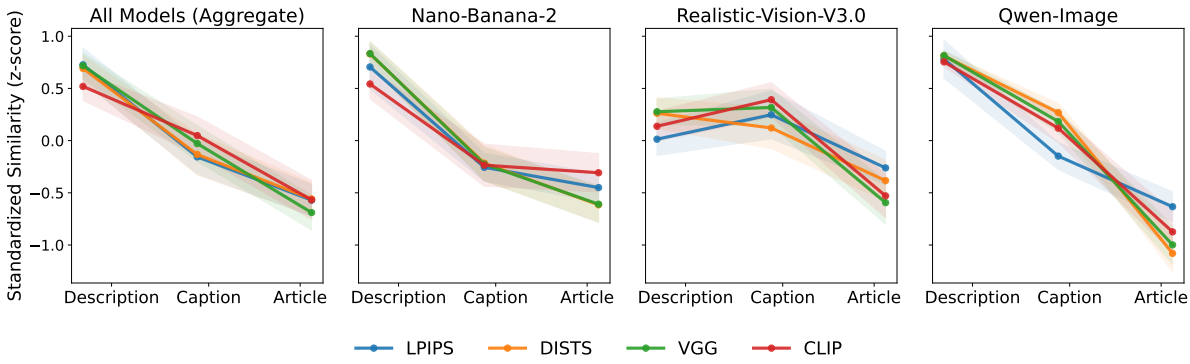


Figure 1: Intra-Prompt Diversity (lower similarity = greater diversity). Metrics are $z$-scored across 108 instances per condition; ribbons denote 95% CIs; results are shown per model and in aggregate (across candidates and models).

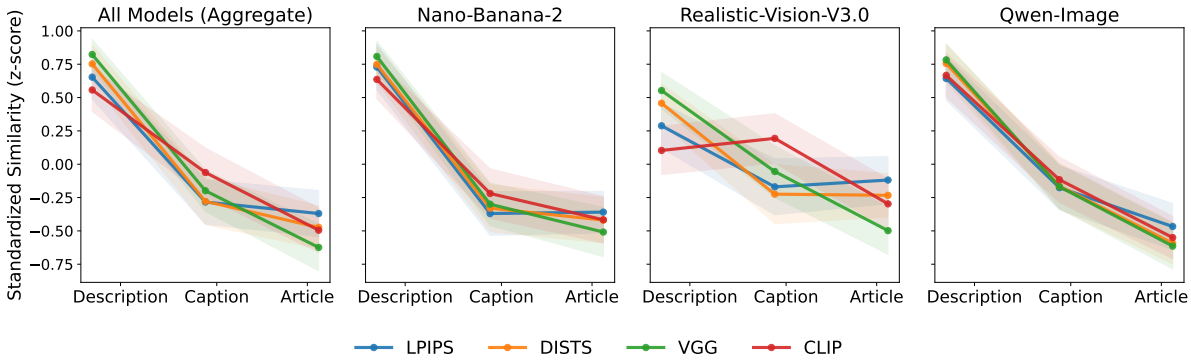


Figure 2: Reference Deviation ((lower similarity = greater devaition). Metrics are $z$-scored across 108 instances per condition; ribbons denote 95% CIs; results shown per model and aggregated (across candidates and models).

perceive and prefer generated imagery.

All procedures were approved by the SCHOOL_NAME Institutional Review Board (IRB Protocol #24-0019-AM-001; May 28, 2025). Experiments were preregistered on AsPredicted (https://aspredicted.org/8kd9ss.pdf).

### 5.1 Preferences Across the VNF Latitude

For this experiment, we switch to a state-of-the-art proprietary TTI system to generate high-fidelity, photorealistic news images. This allows us to measure participant preferences across TTI tasks with increasing VNF latitude in an ecologically valid news context, while minimizing reliance on low-level stylistic artifacts that weaker generators may introduce.

**Stimuli and Model Selection** All image descriptions were generated via `gpt-4o`, following the same process documented in Section 4.1. We selected `Imagen 3` for its strong ability to produce high-fidelity, photorealistic images of coherent human figures and naturalistic scenes. We further conducted extensive controls and validation, ensuring balanced image fidelity (see Appendix D) and confirming that reference deviation increases from description- to article-based generation under `Imagen 3` as textual constraints are relaxed (see Appendix C).

**Procedure** We employed a within-subject design comparing description-, caption-, and article-based generation conditions using a two-alternative forced choice paradigm. On each trial, participants were shown a news article alongside two images: one original journalist-selected image and one generated by a TTI model. Participants selected the image they would publish as a journalist for the given article. The forced-choice design enables direct comparison between TTI-generated and original images under a shared article context, yielding a relative news image preference measure that can be directly interpreted against the 50% chance level.

Each participant viewed images generated under all three conditions but saw each article only once. Participants completed 108 randomized trials (i.e., 36 per condition), with stimuli fully balanced across the three conditions. To encourage adequate


Figure 3: Preference experiment across VNF latitude. Left: example description, caption, and article for a news instance; Middle: real image and generated images under each condition; Right: violin plots of results over 108 stimuli (gray dots), with means and bootstrapped 95% CIs.

engagement with the article content, each trial first presented the article alone with a suggested minimum viewing time of approximately 30 seconds. Participants then proceeded to a second page displaying both the article and its corresponding image for coherence assessment. Two practice trials preceded the main task.

**Participants and Exclusions** 110 participants were recruited from the SCHOOL_NAME student subject pool in January 2026, with no overlap with other studies in this paper. After excluding participants who failed more than one attention check or provided incomplete responses, 98 participants remained for analysis.

**Results** Given the forced choice setup (where 0.50 reflects equal preference between TTI-generated and real images), one-sample $t$-tests showed that both the description- ($M = 0.441$, $p < .001$) and caption-based ($M = 0.435$, $p < .001$) conditions were significantly below chance, indicating a preference for real images. In contrast, the article-based condition was significantly above chance ($M = 0.531$, $t = 2.65$, $p = .009$), indicating a preference for AI-generated images.

We analyzed differences across generation conditions using a linear mixed-effects model with condition as a fixed effect and random intercepts for subjects. The article-based condition was significantly higher than the description-based condition ($\beta = 0.091$, $SE = 0.012$, $z = 7.60$, $p < .001$), indicating that images generated from article-based prompts were more likely to be preferred over real images. Additionally, the caption-based condition did not differ from the description-based condition ($\beta = -0.006$, $SE = 0.012$, $z = -0.50$, $p = .618$), suggesting similar preference levels under more constrained TTI tasks.

## 5.2 Preferences for AI Visual Styles

To examine whether observed stylistic regularities reflect `Imagen 3`-specific behavior or more general properties of contemporary TTI systems, we replicated the description-based generation condition using `GPT-5`–generated descriptions and additional TTI models.

**Stimuli and Model Selection** We synthesized images with three state-of-the-art TTI models: `Imagen 4`, `GPT-Image-1`, and `Qwen-Image`. This design tests whether the stylistic claim persists across model families with different training paradigms, architectures, and stylistic priors. These models are selected to capture diversity across proprietary and open-weight systems. `Imagen 4` is the latest iteration of the Imagen family and serves as a within-family replication of `Imagen 3` (Saharia et al., 2022). `GPT-Image-1`, OpenAI's flagship image generation model, represents a distinct proprietary system with its own learned visual aesthetics and rendering characteristics, enabling stylistic comparison beyond the Imagen lineage. `Qwen-Image` is implemented following the same procedure described in Section 4.2, with two modifications: the number of inference steps is increased to 50, and `true_cfg_scale` is set to 4.0 to improve generation fidelity.

**Procedure** We employed a within-subject design comparing description-based generation across three TTI models, `Imagen 4`, `GPT-Image-1`, and `Qwen-Image`, using the same forced-choice paradigm. In each trial, participants were shown a news article alongside two images: the original image and one generated by one of the three models.

Each participant viewed images generated by all three models but saw each article only once. Participants completed 108 randomized trials (i.e., 36 per model), following the same procedure as in


Figure 4: Preference experiment for AI visual style. Left: the same description used across all generators; Middle: real image and model-generated images; Right: violin plots of results over 108 stimuli (gray dots), with means and bootstrapped 95% CIs.

the previous experiment. All remaining procedures mirrored the prior study.

**Participants and Exclusions** A total of 110 participants were recruited from the SCHOOL_NAME student subject pool between February and March 2026, with no overlap with other studies reported in this paper. After exclusions, 97 participants remained for analysis.

**Results** One-sample $t$-tests showed that all models were significantly below chance (`Imagen 4`: $M = 0.290$, $p < .001$; `GPT-Image-1`: $M = 0.181$, $p < .001$; `Qwen-Image`: $M = 0.180$, $p < .001$), indicating a consistent preference for journalist-selected images over generated images. This highlights a systematic stylistic disadvantage of TTI-generated images under tightly constrained description-based prompts.

Further analyzing model differences using a linear mixed-effects model with random intercepts for subjects, both `GPT-Image-1` ($\beta = -0.109$, $SE = 0.011$, $z = -10.22$, $p < .001$) and `Qwen-Image` ($\beta = -0.110$, $SE = 0.011$, $z = -10.33$, $p < .001$) showed significantly lower preference than `Imagen 4`. This indicates that `Imagen 4` yields comparatively higher-fidelity visual outputs, supporting our use of the Imagen family for high-fidelity image generation in our study.

## 6 Conclusion

We argue that the dominant paradigm in modern VLMs largely assumes tight semantic alignment between images and text, whereas most real-world multimodal contexts exhibit image–text relationships at higher narrative and symbolic levels. This semantic looseness motivates us to examine how different levels of textual constraint shape the Visual Narrative Freedom of Text-to-Image systems. We conduct model experiments across three common prompting settings (description-, caption-, and article-based generation) and introduce two metrics, *Intra-Prompt Diversity* and *Reference Deviation*, to characterize how these conditions reflect VNF. Our results show a gradual increase in VNF as input constraints are relaxed from description- to article-based generation.

Our human evaluation reveals important societal implications. Loosening Visual Narrative Freedom does not only result in generated images that are more preferred than under tighter VNF constraints, it can even lead to these system outputs being preferred over journalist-selected images in high-stakes civic contexts. These findings suggest that the creation of persuasive multimodal misinformation is closer than evaluations focused on TTI systems solely prompted with text descriptions may suggest. These results, therefore, underscore the need to reconsider how current VLM paradigms conceptualize image–text alignment, as the looseness of real-world multimodal relationships, combined with current VLM capabilities, can meaningfully shape content production and perceived persuasiveness.

## Limitations

Due to computational constraints, we restrict VNF measurement to fast-inference TTI models; future work should extend to a broader set of proprietary and open-source generators. Our dataset can be expanded to other multimodal domains (e.g., blog posts (Sharma et al., 2018), Wikipedia (Kreiss et al., 2022b), though this requires additional data collection of articles and images. Finally, analyzing factors such as image–article alignment (e.g., CLIPScore) and AI detectability may further clarify how article-based generation leverages VNF to produce more persuasive content.

## References


Marcus Banks. 2013. True to life: Authenticity and the photographic image. In Thomas Fillitz and Jamie A. Saris, editors, *Debating Authenticity: Concepts of Modernity in Anthropological Perspective*, pages 160–171. Berghahn Books, New York.

Samah Saeed Baraheem, Trung-Nghia Le, and Tam V. Nguyen. 2023. Image synthesis: a review of methods, datasets, evaluation metrics, and future outlook. *Artificial Intelligence Review*, 56(10):10813–10865.

Roland Barthes. 1977. *Image, Music, Text*. Fontana Press, London. Translated by Stephen Heath.

Gullal S. Cheema, Sherzod Hakimov, Eric Müller-Budack, Christian Otto, John A. Bateman, and Ralph Ewerth. 2023. Understanding image-text relations and news values for multimodal news analysis. *Frontiers in Artificial Intelligence*, Volume 6 - 2023.

Massimiliano Ciammaichella. 2023. This person does not exist. representation theories and practices of a desired face. In *Proceedings of the 3rd International and Interdisciplinary Conference on Image and Imagination*, pages 544–551, Cham. Springer International Publishing.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. 2022. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581.

Nanyi Fei, Zhiwu Lu, Yizhao Gao, Guoxing Yang, Yuqi Huo, Jingyuan Wen, Haoyu Lu, Ruihua Song, Xin Gao, Tao Xiang, Hao Sun, and Ji-Rong Wen. 2022. Towards artificial general intelligence via a multimodal foundation model. *Nature Communications*, 13(1):3094.

Yuqi Huo, Manli Zhang, Guangzhen Liu, Haoyu Lu, Yizhao Gao, Guoxing Yang, Jing Wen, Heng Zhang, Baogui Xu, Weihao Zheng, Zongzheng Xi, Yueqian Yang, Anwen Hu, Jinming Zhao, Ruichen Li, Yida Zhao, Liang Zhang, Yuqing Song, Xin Hong, and 16 others. 2021. Wenlan: Bridging vision and language by large-scale multi-modal pre-training. *ArXiv*, abs/2103.06561.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 4904–4916. PMLR.

Elisa Kreiss, Cynthia Bennett, Shayan Hooshmand, Eric Zelikman, Meredith Ringel Morris, and Christopher Potts. 2022a. Context matters for image descriptions for accessibility: Challenges for referenceless evaluation metrics. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4685–4697, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Elisa Kreiss, Fei Fang, Noah Goodman, and Christopher Potts. 2022b. Concadia: Towards image-based text generation with a purpose. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4667–4684, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David Shamma, and 1 others. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Liunian Harold Li, Haoxuan You, Zhecan Wang, Alireza Zareian, Shih-Fu Chang, and Kai-Wei Chang. 2021. Unsupervised vision-and-language pre-training without parallel images and captions. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5339–5350, Online. Association for Computational Linguistics.

Manling Li, Alireza Zareian, Qi Zeng, Spencer Whitehead, Di Lu, Heng Ji, and Shih-Fu Chang. 2020. Cross-media structured common space for multimedia event extraction. In *Proceedings of The 58th Annual Meeting of the Association for Computational Linguistics*.

Songtao Li and Hao Tang. 2026. Multimodal alignment and fusion: A survey. *International Journal of Computer Vision*, 134(3):103.

Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. 2022. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *International Conference on Learning Representations*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*.

Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. 2021. Visual news: Benchmark and challenges in news image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6761–6771, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Minqian Liu, Zhiyang Xu, Zihao Lin, Trevor Ashby, Joy Rimchala, Jiaxin Zhang, and Lifu Huang. 2024. Holistic evaluation for interleaved text-and-image generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22002–22016, Miami, Florida, USA. Association for Computational Linguistics.

Andreas Lommatzsch, Benjamin Kille, Özlem Özgöbek, Yuxiao Zhou, Jelena Tešić, Cláudio Bartolomeu, David Semedo, Lidia Pivovarova, Mingliang Liang, and Martha Larson. 2022. Newsimages: addressing the depiction gap with an online news dataset for text-image rematching. In *Proceedings of the 13th ACM Multimedia Systems Conference*, MMSys '22, page 227–233, New York, NY, USA. Association for Computing Machinery.

Mehmet I. Mehmet and Rodney J. Clarke. 2016. B2b social media semantics: Analysing multimodal online meanings in marketing conversations. *Industrial Marketing Management*, 54:92–106.

Nelleke Oostdijk, Hans van Halteren, Erkan Başar, and Martha Larson. 2020. The connection between the text and images of news articles: New insights for multimedia analysis. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4343–4351, Marseille, France. European Language Resources Association.

Sejin Paik, Sarah Bonna, Ekaterina Novozhilova, Ge Gao, Jongin Kim, Derry Wijaya, and Margrit Betke. 2023. The affective nature of ai-generated news images: Impact on visual journalism. In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR.

Sandra Ristovska. 2020. The need for visual information policy. *Surveillance & Society*, 18(3):418–421.

Daniel Rubinstein and Katrina Sluis. 2008. A life more photographic: Mapping the networked image. *Photographies*, 1(1):9–28.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Kamyar Ghasemipour, Raphael Gontijo-Lopes, Burcu Ayan, S. Sara Mahdavi, and 1 others. 2022. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Annual Meeting of the Association for Computational Linguistics*.

Hyunwoo Shin, Nils Gessert, Wieland Brendel, Simon Wiesler, and Damian Borth. 2022. Cloob: Modern hopfield networks with infoloob outperform clip. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS)*, NeurIPS '22, pages 1487–1505, Red Hook, NY, USA. Curran Associates Inc.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Association for Computing Machinery.

Edina Strikovic and Hannes Cools. 2025. Reality reimag(in)ed. mapping publics' perceptions and evaluations of ai-generated images in news contexts. *Digital Journalism*, 0(0):1–22.

T. J. Thomson, Ryan J. Thomas, and Phoebe Matich. 2025. Generative visual ai in news organizations: Challenges, opportunities, perceptions, and policies. *Digital Journalism*, 13(10):1693–1714.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. In *Transactions of the Association for Computational Linguistics*, volume 2, pages 67–78.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2018. From recognition to cognition: Visual commonsense reasoning. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6713–6724.

Beichen Zhang, Pan Xu, Jiabang Chen, Yutong Zhou, Yu Qiao, and Zheng Shou. 2024. Long-clip: Unlocking the long-text capability of clip. In *European Conference on Computer Vision (ECCV)*. Springer.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595.

Yongshuo Zong, Oisin Mac Aodha, and Timothy Hospedales. 2023. Self-supervised multimodal learning: A survey. *arXiv preprint arXiv:2304.01008*.

Rolf A. Zwaan. 2014. Embodiment and language comprehension: reframing the discussion. *Trends in Cognitive Sciences*, 18(5):229–234.

## A Descriptive Statistics of News Datasets

We report descriptive statistics of our news dataset in Table 1, showing topic distributions that broadly reflect the underlying outlets and major news categories, along with average token counts by topic. We observe no substantial differences in length across topics, and all textual inputs fall within the LongCLIP context window (i.e., 248 tokens).

| News Topics | N | Description | Caption | Article |
|---|---|---|---|---|
| Politics & International Affairs | 56 | 217.2 | 28.1 | 229.0 |
| Science, Technology & Environment | 20 | 200.4 | 30.6 | 222.8 |
| Culture & Lifestyle | 15 | 199.7 | 28.7 | 215.0 |
| Sports | 11 | 220.0 | 25.6 | 210.0 |
| Economy & Business | 6 | 214.7 | 25.8 | 226.2 |
| **Total/Average** | **108** | **210.4** | **27.8** | **220.6** |

Table 1: Distribution of news topics and average token lengths for article text, captions, and generated image descriptions in the dataset. Token counts follow the LongCLIP tokenizer (CLIP BPE based on the GPT-2 vocabulary; max context length 248 tokens).

## B Generation Prompts

**Image Description Generation (GPT-4o/GPT-5)**

> Describe the image in a way that could serve as a photorealistic news image prompt for an AI model. Focus on the following elements: the subjects (their appearance, approximate number, and actions), the environment (location and background), the mood (emotional tone), the lighting (source, direction, and intensity), shadows, and realistic details (textures, clothing, facial expressions). Provide a realistic estimate of the number of visible people or subjects, avoiding exaggeration. Also describe the camera angle or perspective (e.g., high angle, low angle, over-the-shoulder). Write in a descriptive, journalistic tone as a single uninterrupted paragraph. Do not include labels, bullet points, or line breaks.

**TTI Prompting Conditions** For each news instance, we generated images under three conditions using aligned prompts derived from the same underlying content:

- **Description-based:** "Generate a photorealistic news image based on the following description."
- **Caption-based:** "Generate a photorealistic news image based on the following caption."
- **Article-based:** "Generate a photorealistic news image representing the main subject of the following article. No text, headlines, or labels in the image."

The explicit prohibition of textual overlays in the article-based condition was necessary because TTI systems frequently embed article text directly within images when under-instructed.

## C Reference Deviation of `Imagen 3`

For the high-fidelity image model `Imagen 3` used in the preference experiment across VNF latitude, we validate that reference deviation increases (i.e., similarity decreases) progressively from description- to caption- to article-based generation across perceptual, structural, semantic, and conceptual measures. This pattern indicates that, as textual constraints are relaxed, `Imagen 3` generates images that increasingly diverge from journalist-selected images across visual representations, thereby expanding its capacity to leverage VNF and potentially influencing audience image preferences.

## D Additional Details on Balancing Image Fidelity

To control for perceptual confounds between original and TTI-generated images, we apply extensive preprocessing steps.

First, all images are center-cropped to match aspect ratios when TTI outputs do not precisely align with the originals.

Second, because archival news images often appear at relatively low resolution—a salient cue of authenticity—we manually retrieve the highest-resolution publicly available versions via Google Image Search. We avoid applying any post hoc resolution enhancement, preserving realistic visual differences between circulated news images (often lower resolution) and generated images (typically higher resolution).

Third, to reduce residual resolution disparities, we apply controlled JPEG compression to both original and generated images. Compression is iteratively applied (quality = 0.9 per iteration) until file sizes fall below 200KB, ensuring consistent loading performance in Qualtrics while maintaining perceptual comparability.

All stimuli are manually inspected by the re-

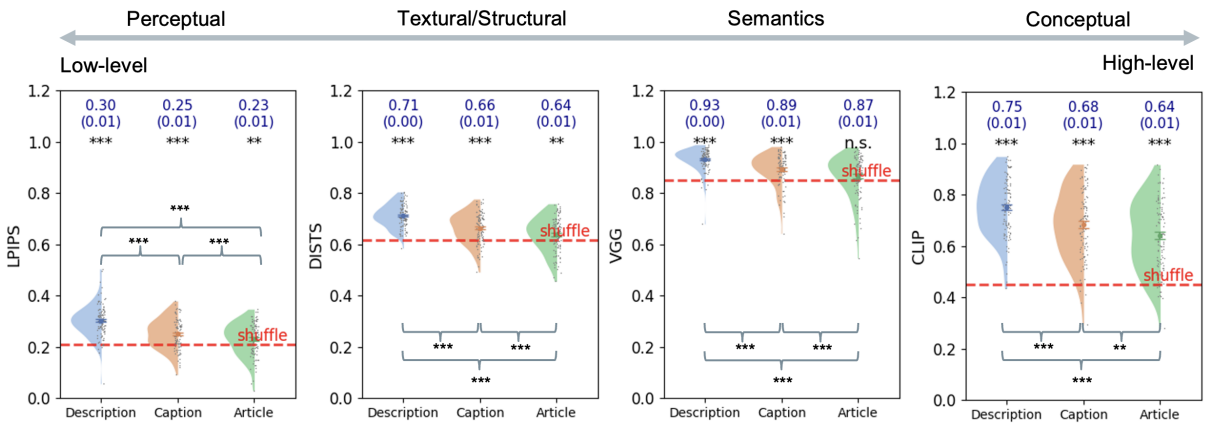


Figure 5: Reference deviation of `Imagen 3`-generated images across three generation conditions, spanning lower-level perceptual to higher-level conceptual similarity. Half-violins show distributions over 108 stimuli (gray dots), with means and bootstrapped 95% CIs. Lower similarity indicates greater deviation from the original image. The shuffled baseline (red dashed line) reflects similarity between original images and randomly paired images, serving as a lower-bound reference. Significance markers ($^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$) denote comparisons to the original (top) and between conditions (bottom).

search team to ensure no salient resolution differences remain beyond inherent stylistic variation. Implausible or nonfactual generations (e.g., a floating truck in the ocean) are excluded or regenerated prior to inclusion. These refinements help minimize superficial perceptual cues that participants might otherwise rely on when assessing preferences between real and generated images.

The same procedures are applied across both human experiments in Section 5.