RatioMorph: Controllable Diffusion Framework for Automotive Viewpoint and Proportion Manipulation in Vehicle Design

Haeji Go[*1], Jae-Hun Lee[*1], Shinyeong Noh[*1], Kayoung Kim[*1], Kyuseong Lim[*1], Jee Eun Song[*1†], Mingyu Lee[*1], Joowan Sung[2], Soonbeom Kwon[2], Myoungbok Shin[2], Junsang Park[2]

[1]LG CNS AI Center
[2]Hyundai Motor Company
{haejigo, Jaehun.Lee, sy.noh, gy.kim, ks.lim, jeeeun.song, cinchon}@lgcns.com
{joowansung, sixthfinger, 7139975, junsang.park}@hyundai.com

Abstract

Designing vehicle exteriors requires repeated refinement of key proportions and viewpoints, a process traditionally reliant on manual sketching, which is often time-consuming and inefficient in early concept stages. To accelerate the design process, we are exploring the potential of utilizing AI for ideation in these early stages. However, it remains a challenging task to control proportions and maintain a fixed perspective when generating images using AI. To address these limitations, we present RatioMorph, a controllable image generation system that enables manipulation of vehicle proportions and viewpoints when generating images by AI. RatioMorph comprises two core modules. Car2BoxNet is a depth estimation model that transforms real photographs into structured box-style depth maps that capture the geometric layout of the vehicle. Box2CarNet is a diffusion-based image generator fine-tuned to produce vehicle designs that adhere to the provided geometric conditions. Both Car2BoxNet and Box2CarNet are trained on a synthetic dataset curated through automated filtering based on geometric alignment and visual quality. Evaluated within a production-adjacent automotive design workflow, RatioMorph significantly reduced early-stage design iteration time and enabled exploratory workflows that were difficult with previous AI workflows. This work introduces a domain-specific, controllable diffusion-based generation system tailored for automotive design, enabling manipulation of vehicle viewpoint and proportion. It demonstrates strong potential to accelerate early-stage workflows and outlines a path toward industrial deployment, with phased integration into production environments currently underway.

Introduction

The exterior design of an automobile defines the visual identity of a vehicle and significantly affects consumer perception. Among visual attributes, viewpoint and proportion control—referring to the relative scale and placement of components such as the wheelbase, cabin, and overhangs—play a central role in achieving aesthetic balance and maintaining brand consistency. Existing computational tools offer limited support, particularly for generating proportionally varied concepts under fixed viewpoints, a common constraint in early ideation stages. As automotive markets evolve rapidly and customer preferences shift, reducing development time has become a critical challenge. Efficient exploration of proportionally diverse design concepts is essential. This motivates the need for intelligent systems that assist designers during early-phase ideation, while preserving creative control and design feasibility.

Traditionally, vehicle development follows a multi-stage design workflow, beginning with the ideation of multiple sketches across various angles (front-quarter, side, rear-quarter views), followed by internal reviews, selection of promising concepts, 3D modeling, and final rendering. In global automotive companies, dozens of designers collaborate and compete during these early stages to define a final production-ready form, often iterating rapidly within tight timelines. While this iterative process fosters creativity, it also introduces inefficiencies and redundant manual work, especially during the ideation phase, where designers explore countless variations in form and proportion. Recognizing this bottleneck, we identify the ideation stage as a key opportunity where AI can provide meaningful support by generating diverse design proposals with controlled proportions from fixed viewpoints, thus streamlining concept exploration.

To address the aforementioned challenges, we present RatioMorph, an AI-assisted design support framework that enables realistic and controllable vehicle image generation under fixed viewpoints. The framework integrates two components. First, a generative module composed of Car2BoxNet and Box2CarNet enables fine-grained control over vehicle proportions while maintaining stylistic coherence. Car2BoxNet is a depth estimator trained to infer structured box-style geometry from real vehicle images, and Box2CarNet is a controllable diffusion model that synthesizes photo-realistic vehicle designs conditioned on these depth maps and text prompts. Second, RatioMorph provides

[*]These authors contributed equally.
[†]Corresponding author. Email: jeeeun.song@lgcns.com

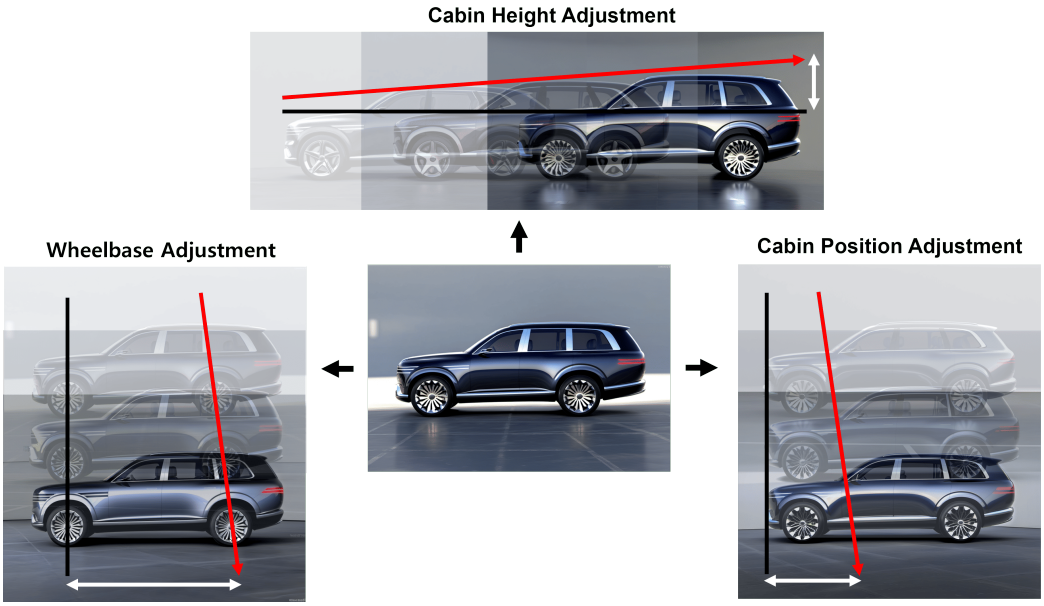

Figure 1: Core capabilities of RatioMorph: controllable manipulation of vehicle proportions from a single image. The system enables adjustment of key design parameters, including wheelbase (left), cabin height (top), and cabin position (right), enabling rapid visual exploration in early-stage concept development.

an interactive design workflow that accepts numerical parameters, 3D box sketches, and real vehicle photos, converting them into structured depth conditions for generation. These serve as inputs to the generation pipeline, allowing designers to iterate on vehicle concepts with minimal manual effort. This accelerates rapid early-stage ideation, empowering designers to focus on creative direction and proportion exploration.

To summarize, our work addresses a critical gap in automotive design automation by enabling precise and controllable proportion manipulation under fixed viewpoints. Through the integration of domain-specific data and a novel generative architecture, we provide a practical tool that enhances early-stage ideation efficiency and supports creative exploration. Our key contributions are as follows.

- We introduce a novel AI framework that enables proportion-controllable image generation for automotive design ideation under fixed viewpoints.
- Our method supports designers in the early stages of concept development by expanding the breadth of visual exploration with minimal manual effort.
- Through both qualitative and quantitative evaluation, we demonstrate the limitations of existing generative models in proportion manipulation, and validate that our system reduces design iteration time while preserving visual fidelity.

Related Works

Our work focuses on combining efficient diffusion fine-tuning with explicit spatial control and monocular depth estimation for geometric conditioning.

Efficient Diffusion Adaptation with Spatial Conditioning

Parameter efficient adapters such as LoRA Adapter (Shih et al. 2023) employ low-rank factorization to fine-tune large text-to-image diffusion models with minimal overhead. Spatial conditioning methods extend this paradigm by injecting geometry signals into the diffusion process: ControlNet (Zhang and Agrawala 2023) and LooseControl (Bhat, Mitra, and Wonka 2024) introduce depth- and box-based conditioning, while cross-attention-based pipelines enforce user-specified layouts. Complementary advances include SDXL (Podell et al. 2023), which scales latent diffusion to high resolutions, and recent vision language models (Lu et al. 2019; Radford et al. 2021), which provide rich textual conditioning from images. These works collectively demonstrate that lightweight adaptation plus explicit spatial control yields both fidelity and flexibility in generative tasks.

Depth Estimation and Geometric Conditioning

Monocular depth estimation is fundamental for enforcing geometric constraints in image synthesis. Multiscale CNNs (Eigen, Puhrsch, and Fergus 2014) established the basis for predicting the depth of a single image and downstream filtering, while transformer-based models such as DPT (Ranftl, Bochkovskiy, and Koltun 2021) improve geometric detail recovery. LooseControl (Bhat, Mitra, and Wonka 2024) integrates these depth predictions to guide diffusion outputs toward accurate object outlines. Building on this foundation, our Box2CarNet achieves precise depth-box localization, resulting in significantly tighter geometric alignment in generated images.

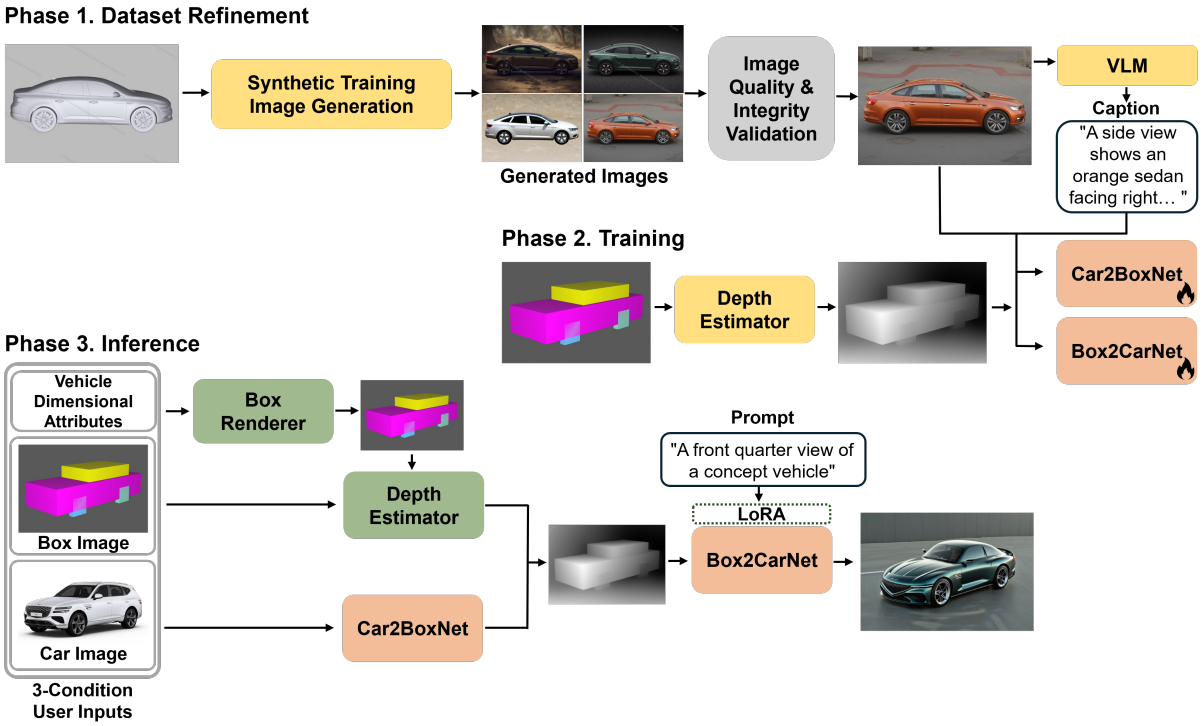

Figure 2: Overview of the RatioMorph framework. The system consists of three phases: (1) *Dataset Refinement*, which generates and filters synthetic training images with quality and alignment checks; (2) *Training*, where Car2BoxNet and Box2CarNet are trained using structured depth conditions; and (3) *Inference*, which supports three user input types (dimensional attributes, box images, or photos) to produce high-fidelity vehicle images through controllable generation.

Framework : RatioMorph

Our framework, *RatioMorph*, is a comprehensive system designed to translate a designer's high-level intent into photo-realistic and structurally consistent vehicle images. The framework operates in three stages: (1) a dataset construction pipeline that builds a high-quality foundation, (2) model training procedures for both condition generation and image synthesis, and (3) an inference framework that supports diverse user inputs and enables flexible, controllable image generation.

Phase 1: Dataset Construction and Filtering

Existing datasets such as KITTI (Geiger et al. 2013) and Cityscapes (Cordts et al. 2016) are unsuitable for design-oriented generation, as they focus on perception tasks and contain images with multiple objects, inconsistent viewpoints, and insufficient geometric annotations. To address this, we construct a large-scale synthetic dataset tailored for vehicle design generation.

Synthetic Data Generation We render diverse 3D vehicle models from a wide range of controlled camera angles. For each rendered image, we prepare a corresponding 3D box condition map composed of the vehicle's main components such as wheels, the cabin, and the body. These box images reflect the core geometry and spatial layout of the vehicle. To generate photo-realistic appearance images consistent with the given structural constraints, we leverage a pre-trained conditional diffusion model (Ho, Jain, and Abbeel 2020; Dhariwal and Nichol 2021). The resulting images are then enhanced using deep learning-based super-resolution models (Li et al. 2022; Ledig et al. 2017) to improve visual fidelity and resolution, producing high-quality samples. Each synthesized image is paired with a caption generated using a vision-language model (VLM).

Automated Data Filtering To ensure alignment between the vehicle image and its box condition, we compute a mask-based misalignment score:

$$S = \frac{A_{\text{box-only}}}{A_{\text{box-total}}} + \frac{A_{\text{vehicle-only}}}{A_{\text{vehicle-total}}} \quad (1)$$

where $A_{\text{box-only}}$ and $A_{\text{vehicle-only}}$ are the non-overlapping regions of the box and vehicle masks, respectively. Samples with a score $S > \tau$ are discarded. Additionally, we apply aesthetic filtering using a pre-trained image quality assessment model (Wang et al. 2004; Talebi and Milanfar 2018) to retain samples of higher visual appeal.

Phase 2: Model Training

Car2BoxNet Car2BoxNet is a task-specific depth estimation model trained to infer abstract 3D box-style depth maps from real-world vehicle images. While synthetic data offers paired images and geometry, real images typically lack

structural annotations. To address this, we trained a depth estimation model (Eigen, Puhrsch, and Fergus 2014) using our synthetic dataset, allowing Car2BoxNet to generalize from real images. The output depth maps preserve the core structural layout of the vehicle in a simplified form, enabling consistent geometric conditioning during generation. Once trained, Car2BoxNet is used to expand training coverage and support real-image inputs during inference.

Box2CarNet Box2CarNet is a fine-tuned ControlNet model designed to generate photo-realistic vehicle images guided by structured depth maps and textual descriptions. Unlike conventional depth conditioning, our method uses structured 3D box-style depth maps, enabling fine-grained control over part-specific attributes such as wheel placement, cabin size, and body proportions. Box2CarNet is trained using triplets of a 3D box-style depth map, caption, and target image, where the depth map is either derived from synthetic box images or predicted by Car2BoxNet. The model learns to synthesize realistic designs that are both geometrically aligned and semantically consistent with the input prompt.

Phase 3: Inference Framework

Processing of Multiple Input Types The inference system supports three input types—numerical attributes, 3D box images, and real vehicle photos—all converted into box-style depth maps serving as geometric conditions for final image generation. Numerical attributes specify vehicle dimensions and camera angles to construct a 3D box representation. 3D box images are transformed into depth maps via standard estimation, while real photos are processed by Car2BoxNet to infer corresponding depth maps capturing vehicle structure.

Controllable Generation with Viewpoint and Proportion Once the structured depth map is obtained, it is combined with a user-provided text prompt and fed into Box2CarNet to generate high-resolution vehicle images that are both photo-realistic and structurally consistent. Box2CarNet uses the 3D box-style depth map to precisely control key components—such as wheels, cabin, and body—while following the specified viewpoint and proportions. Interpretable parameters encoded in the depth map enable part-aware, geometry-consistent synthesis that aligns with both stylistic intent and structural constraints. The vehicle's geometric structure is defined by key proportion parameters—such as body length, height, and wheelbase—which determine the relative placement and scale of individual parts within the box representation. For viewpoint control, the system allows flexible adjustment of both horizontal and vertical camera angles. Together, these settings allow for part-aware generation with precise control over both structural proportions and viewing perspective.

Stylistic Customization To support lightweight customization without compromising structural controllability, our framework can integrate LoRA-based modules during inference. These modules enable users to adapt the generation process to specific stylistic domains, brand identities, or design cues without retraining the entire model. This allows designers to rapidly explore diverse styles while preserving the defined viewpoint and proportions.

Evaluation

In this section, we present a comprehensive evaluation of our proposed model against the conventional LooseControl(Bhat, Mitra, and Wonka 2024) baseline. Our evaluation protocol comprises three components: (i) quantitative comparison between the LooseControl model and our method using the geometric alignment filtering score S in Equation (1), (ii) quantitative assessment of task duration, and (iii) qualitative visual presentation of diverse inference outputs to illustrate the practical advantages of our method.

Comparison with LooseControl Baseline

We evaluate our method against the conventional LooseControl (LC) baseline by measuring the geometric misalignment score S defined in Equation (1). We compute the misalignment score S for both models and compare the average and variance of misaligned-pixel proportions to quantify improvements in geometric alignment. Table 1 reports S for both approaches across multiple vehicle types and views. Our method consistently achieves lower misalignment than LC, indicating tighter adherence to prescribed bounding-box constraints. Figure 3 provides a side-by-side comparison of the outputs of LC and our method on representative cases. While the LC baseline often produces noticeable shape distortions and background bleed-through around the vehicle edges, our method preserves sharper contours and more accurate geometry, validating the quantitative gains observed in Table 1.

Vehicle Type	View	S_{LC}	S_{Ours}
Sedan	front quarter view	**0.094**	0.219
Sedan	front view	0.196	**0.158**
Sedan	side view	0.233	**0.173**
SUV	front quarter view	**0.085**	0.227
SUV	front view	0.193	**0.176**
SUV	side view	0.321	**0.195**

Table 1: Quantitative comparison of misalignment scores between loose control (LC) and our method across vehicle types and views, showing consistently lower S values for our approach.

Task Duration Analysis

We quantify human and computational effort by comparing end-to-end design iteration times before and after adopting our system. As these measurements do not reflect proportion and viewpoint adjustments to a fixed design, the reported absolute times should be regarded as reference values only. AI-generated images also require additional designer modifications for refinement, with an average time of approximately two hours. As shown in Table 2, our method achieves substantial reductions in design iteration time across all measured tasks. Specifically, image generation is accelerated

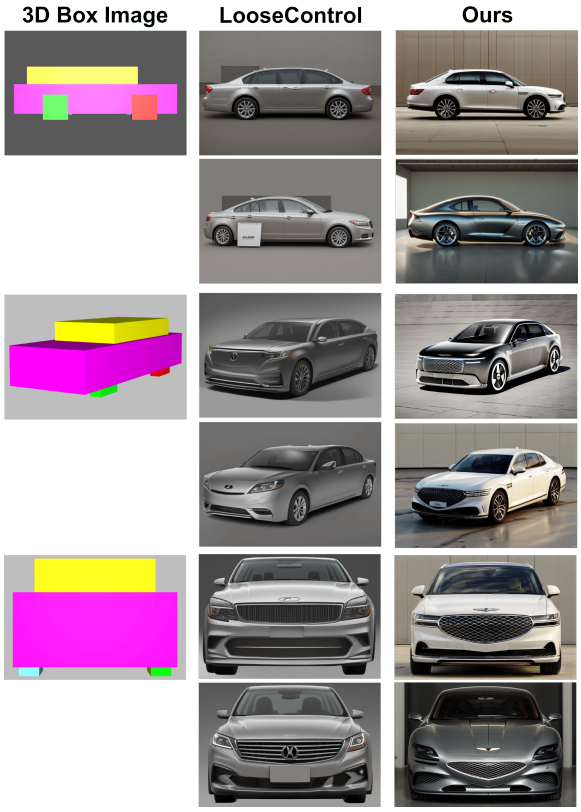

Figure 3: Qualitative comparison between LooseControl and Ours. LooseControl often produces blurred or misaligned boundaries, whereas our method generates sharp contours and accurately conforms to the 3D box constraints across varied vehicle examples.

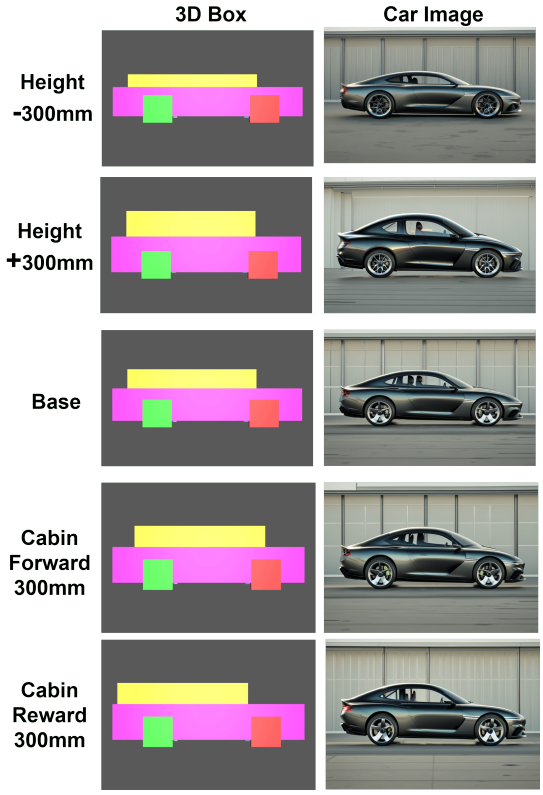

Figure 4: Specification Adjustments. Cabin height changes of ±300 mm and cabin-position shifts of ±300 mm (forward/rearward) within the 3D box produce correspondingly scaled/translated vehicle images.

from manual sketching to prompt-based synthesis, aspect-ratio adjustments are reduced from manual Photoshop editing to brief prompt tuning, and viewpoint framing is similarly reduced from hours of manual work to minutes of prompt-based adjustment. This streamlined workflow enables designers to explore, refine, and iterate on early-stage concepts far more rapidly, while maintaining creative control and ensuring design feasibility.

Inference Results of Our Method

In this section, we present representative inference results and demonstrate how our model yields more coherent and contextually aligned outputs in real-world scenarios. We present qualitative results for parameter-conditioned generation with Box2CarNet and box-style depth maps produced by Car2BoxNet. Figures 4 and 5 show geometric constraints, generated variants under modified parameters, and real-world references, demonstrating precise shape control and photo-realistic fidelity. Figure 6 compares a standard depth estimator with our Car2BoxNet. Car2BoxNet produces vehicle-specific box-style depth maps for direct use as conditioning in our framework.

Path to Deployment

To translate our system into a practical design support tool, we outline a deployment strategy focused on real-world integration and iterative refinement.

Limitations and Future work

While the current model demonstrates potential in generating proportionally controlled vehicle design images, it is trained only on AI-generated 2D data. This leads to limitations in structural consistency, surface details, and viewpoint accuracy, with the lack of physically grounded geometry particularly constraining multi-view coherence and precise proportion control. To address these limitations, we plan to construct a new training dataset by rendering high-resolution images from 3D CAD models. These will be systematically generated to cover diverse viewpoints and parametric proportions, allowing the model to learn from geometrically accurate and richly detailed visual data. This enhancement is expected to improve structural realism and the controllability of both proportion and viewpoint in generated outputs. These enhancements lay the groundwork for on-premises deployment and real-world evaluation, as detailed in the following section.

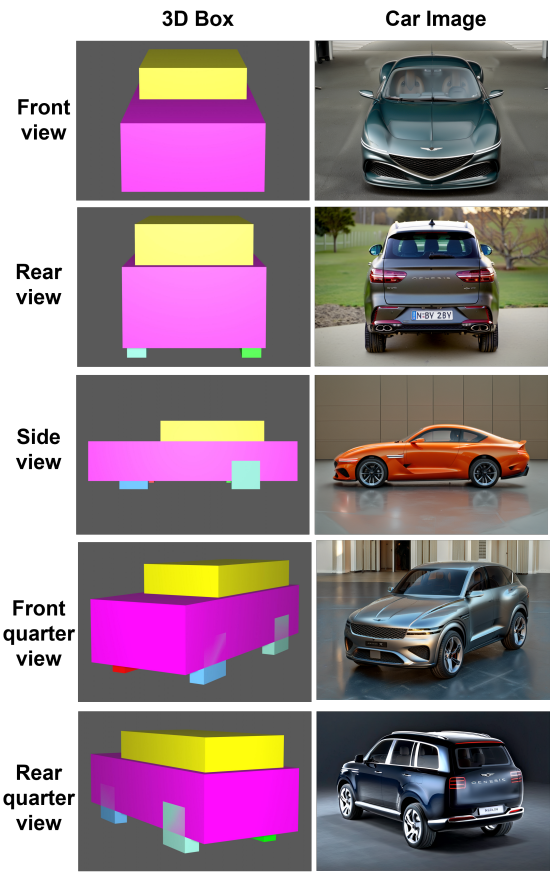

Figure 5: Viewpoint Adjustments. Generation results from multiple views showing consistent geometric compliance and photorealistic fidelity.

Task	**Before (Manual)**
Design generation	8 h (manual sketch)
Aspect-ratio adjustment	4 h (*Photoshop* editing)
Viewpoint framing	8 h
Total Time	**20 h**
Task	**After (RatioMorph)**
Design generation	10 min (incl. prompt creation)
Aspect-ratio adjustment	10 min (incl. ratio input)
Viewpoint framing	10 min (incl. prompt creation)
Total Time	**30 min**

Table 2: Task duration before and after adopting RatioMorph. RatioMorph reduces design iteration time across multiple early-stage tasks.

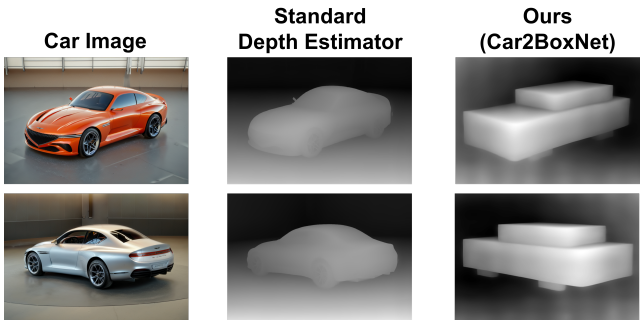

Figure 6: Comparison between a standard depth estimator and Car2BoxNet. Car2BoxNet can produce box-style depth maps tailored to the input vehicle.

Deployment Strategy

Once the new dataset is prepared and the model retrained, we will perform a quantitative evaluation to verify the improvements. The validated model will then be deployed in an on-premise test environment, allowing integration with existing design workflows while maintaining data security and operational alignment.

In the initial phase, the system will be applied to assist designers during early-stage ideation, particularly in generating diverse and realistic design sketches with controlled proportions. Based on internal usage and feedback, we will iteratively improve the model and deployment pipeline. Following successful testing and refinement, we plan a phased rollout across additional design teams and vehicle categories. This gradual deployment strategy will facilitate stable integration and enable technical and organizational adjustments. Ultimately, the system is designed to complement existing workflows during early concept development, allowing the designers to explore proportionally varied sketches more efficiently and with greater flexibility, without disrupting established processes.

This work was conducted as a proof-of-concept (PoC) in collaboration with Hyundai Motor Company. All intellectual property rights arising from this research are owned by Hyundai Motor Company.

Conclusion

We introduced RatioMorph, a controllable diffusion framework for generating proportionally consistent vehicle designs from fixed-viewpoint inputs. The system integrates structured geometric conditioning with a dual-model architecture—Car2BoxNet for depth estimation and Box2CarNet for photo-realistic synthesis—supporting multiple input modalities. Evaluations show improved geometric alignment over baseline models and reduced iteration time, enabling workflows such as architecture transfer and rapid viewpoint exploration in early ideation. While the current model is trained on synthetic data, we plan to build a CAD-based dataset and deploy the enhanced model in on-premise environments. A phased rollout across design teams will support refinement and integration into production workflows. RatioMorph demonstrates a task-specific application of controllable diffusion in industrial design, combining technical feasibility with deployment potential for accelerating early-stage concept development.

References

Bhat, S. F.; Mitra, N.; and Wonka, P. 2024. Loosecontrol: Lifting controlnet for generalized depth conditioning. In *ACM SIGGRAPH 2024 Conference Papers*, 1–11.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

Eigen, D.; Puhrsch, C.; and Fergus, R. 2014. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems (NeurIPS)*.

Geiger, A.; Lenz, P.; Stiller, C.; and Urtasun, R. 2013. Vision meets robotics: The kitti dataset. *The international journal of robotics research*, 32(11): 1231–1237.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4681–4690.

Li, H.; Yang, Y.; Chang, M.; Chen, S.; Feng, H.; Xu, Z.; Li, Q.; and Chen, Y. 2022. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479: 47–59.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Podell, D.; English, Z.; Lacey, K.; Wolters, B.; Tost, G.; Mizrahi, S.; Bitton, E.; Dror, D.; Vinker, Y.; Blattmann, A.; ; et al. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. *arXiv preprint arXiv:2307.01952*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Ranftl, R.; Bochkovskiy, A.; and Koltun, V. 2021. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, 12179–12188.

Shih, Y.-J.; Min, F.; Li, Z.; Liu, S.; and Liu, J. 2023. LoRAdapter: Additive LoRA for Capable and Efficient Text-to-Image Diffusion Fine-Tuning. *arXiv preprint arXiv:2308.05834*.

Talebi, H.; and Milanfar, P. 2018. NIMA: Neural image assessment. *IEEE transactions on image processing*, 27(8): 3998–4011.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Zhang, L.; and Agrawala, M. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*.