Visual Composition Generation of Multi-Source Heterogeneous Concepts – A Practical Study Based on the AIGC Short Film: *The Meeting*

Supplementary Material

Shiqin Hou[1], Jiayan Chen[1], Tony Zhang[1], Baoyang Chen[2], Anyi Rao[1]
[1]MMStudio@HKUST, [2]VisLab@HKUST & Central Academy of Fine Arts
{shouaf, jchenkv, tianyi.zhang}@connect.ust.hk, baoyang.chen@gmail.com, anyirao@ust.hk

A. Theoretical Implications

This experiment yields several observations that may inform ongoing research on concept composition.

Perhaps the most immediate takeaway concerns the relationship between semantic distance and composition difficulty. As our analysis in the main text illustrates, compositing concepts drawn from a shared visual tradition (say, ink-wash landscape with *gongbi* flower-and-bird painting) tends to proceed with manageable friction. Once the concepts span genuinely different visual domains, however, the challenge escalates sharply rather than gradually: pairing traditional ink-wash aesthetics with contemporary architectural photography, for instance, routinely provokes style conflicts that no amount of prompt tuning could fully reconcile. This non-linear scaling hints that the field would benefit from quantitative measures of "inter-concept compatibility" or "semantic distance," which could help predict composition difficulty before generation begins and guide practitioners toward appropriate decomposition or compromise strategies.

A second observation concerns the temporal dimension. Most existing work on concept composition addresses the co-presence of multiple concepts within a single image [3]. In practice, though, video production surfaces a distinct and equally demanding problem: maintaining concept consistency across shots while simultaneously constructing narrative coherence between them. This intermediate level, situated between single-shot frame-to-frame stability and full multi-shot storytelling, remains largely unaddressed. Progress here would be a prerequisite for AI-assisted tools to evolve beyond static image generators into genuinely useful video creation assistants.

Finally, it is worth noting that the workarounds creators arrive at through trial and error often mirror, in structural terms, solutions proposed independently by the research community. Decomposing a compound concept along its sub-dimensions follows a logic closely related to strategies that break complex generation conditions into separately optimizable sub-tasks [1, 3]. The layered approach to spatial consistency control shares its premise with layout-guided [4] and conditional control generation [7]. And the practice of structured reconstruction to counteract cumulative degradation converges, from a purely practical direction, on problem formulations familiar from diffusion-based editing research [5, 6]. These parallels are arguably not accidental, since concept composition itself imposes structural constraints (attributes need decoupling, spatial layout needs explicit representation, iterative refinement needs fault tolerance). If so, systematically documenting the strategies that emerge from creative practice could offer technical research a complementary lens to define problems and set priorities.

B. Limitations and Future Directions

Several limitations should be acknowledged. This is a single-case study in which the same person serves as both creator and analyst, so the generalizability of the findings and the objectivity of the assessments inevitably warrant caution. We have tried to mitigate this by separating subjective judgments from verifiable technical observations wherever possible and by stating evaluative criteria explicitly. A further constraint is that all tools used are commercial, consumer-grade products whose model architectures and training data remain opaque; this limits the precision with which we can attribute observed phenomena to specific technical causes. In addition, the quality assessment of the finished film relied mainly on the creator's own judgment, without systematic quantitative metrics [3].

Several directions emerge as worth pursuing. On the tooling side, embedding explicit spatial modeling capabilities [2, 7], scene-level semantic segmentation, and inter-shot narrative logic into consumer-grade creative software would go a long way toward improving end-to-end efficiency in AI-assisted video production. On the methodological side, developing lightweight pre-assessment mechanisms for inter-concept compatibility could save substantial trial-and-error effort in practice. On the evaluation side, building a systematic quality framework for multi-concept compositional video that accounts for concept fidelity, style consistency, spatiotemporal coherence, and narrative completeness at once remains both open and pressing.

References

[1] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM transactions on Graphics (TOG)*, 42(4):1–10, 2023. 2

[2] Lukas Höllein, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. Text2room: Extracting textured 3d meshes from 2d text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7909–7920, 2023. 2

[3] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023. 2

[4] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22511–22521, 2023. 2

[5] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 2

[6] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6038–6047, 2023. 2

[7] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 2