Appendix

A. Related Work

Compositional generalization. Compositional generalization has garnered significant attention from the generative model community, especially for text-to-image generation. One avenue explores fine-tuning text-to-image models by incorporating feedback from image understanding systems as a form of reward [15, 27, 61, 73]. However, this strategy may be limited by the text comprehension capabilities of models like CLIP. Another approach involves adjusting the models' cross-attention mechanisms to align with the spatial and semantic details specified in the prompts [3, 8, 10, 16, 32, 40, 42, 53, 72]. This approach relies on the interpretability of the foundational models and often results in only broad, suboptimal control over the generated images. By leveraging the planning and reasoning strengths of language models, researchers have also broken down complex prompts into multiple regional descriptions, providing more precise conditions to guide the image generation process [11, 17, 18, 41, 66, 75]. This decomposition aids in creating images that more accurately reflect the detailed components of the prompts. These methods operate at the inference time and do not fundamentally learn disentangled concepts. Recent work [24, 69] utilizes diffusion timesteps to modify the text embedding for refined generation control. Nevertheless, Hu et al. [24] do not consider the spatial relations among concepts and fully rely on the pre-trained diffusion model's capacity. Wu et al. [69] introduce additional inference-time optimization overhead and depend on the CLIP score as the optimization objective, which limits the generation quality with CLIP's capacity. Guided by our theoretical insights, our work imposes proper constraints and modifications on the cross-entropy to learn disentangled concepts and their relations.

Although empirical studies are abundant in the field [3, 14, 24, 26, 42, 75, 83], theoretical understanding remains limited and often hinges on restrictive assumptions about concept interactions. While recent generative [43] and discriminative [63, 64] models show compositional structures can emerge with scale, our work seeks to formally characterize the underlying causal data structures that guarantee such generalization. From the theoretical standpoint, Brady et al. [6], Wiedemer et al. [67] consider concepts that affect disjoint pixel regions, effectively eliminating interaction between them. Lachapelle et al. [38] models the influences of concepts on the pixel space as purely additive, an approach that Brady et al. [7] extends to include second-order polynomial terms. Additionally, Wiedemer et al. [68] assumes direct access to the function governing concept interactions. These theoretical works also tend to overlook the varying levels of abstraction among concepts and their relationships within the latent space. In contrast, thanks to the hierarchical structure, our theory admits compositions of transformations across hierarchical levels, allowing for complex interaction among concepts at different hierarchical levels. Our sparsity regularizer ($\mathcal{L}_{\mathrm{n}}$ in Eq. 6) is related to the "interaction asymmetry" regularizer of, which also penalizes attention overlap, though Brady et al. [7] use a pixel-wise product loss for VAE latent slots rather than a set-based DICE loss for time-dependent, text-conditioning in diffusion.

Latent hierarchical model identification. Modeling complex real-world data requires capturing hierarchical structures among latent variables. Prior work has explored identification conditions for such hierarchies with continuous latent variables influencing each other linearly [1, 13, 25, 71]. Other studies focus on fully discrete cases, limiting their applicability to continuous data like images [12, 21, 45, 81]. Moreover, latent tree models connect variables through a single undirected path [12, 45, 81], which may oversimplify complex relationships. Closely related to ours, Kong et al. [36] address nonlinear, continuous latent hierarchical models. However, their framework cannot identify latent variables component-wise, leaving room for concept entanglement. In contrast, we provide component-wise identifiability for latent variables and the graphical structures, along with transparent conditions for the data-generating process.

Broader connections. Our work shares motivations with several research directions aiming at reasoning beyond fixed supports. For instance,Open-world estimation [4, 56] builds classifiers robust to unseen classes; probabilistic databases [59] handle incomplete data; and hierarchical topic models [5, 48] discover multi-level thematic structure in text corpora. Our technical focus differs from these works: we target *compositional generation* in the image space via an identifiable hierarchical structure with continuous latent variables, as opposed to detecting unknowns (open-world estimation), querying incomplete relations (probabilistic databases), or performing bag-of-words inference with discrete topics (hierarchical topic models).

B. Proofs for Theoretical Results

B.1. Proof for Theorem 3.1

Theorem 3.1 (Composition Generalization)**.** *We assume the data-generating process equation 1. The discrete concept combination* $\mathbf{d}$ *is composable (i.e.,* $\mathbf{d} \in \Omega_{\mathrm{comp}}$*) if for each continuous latent variable* $z \in \mathbf{z}$*, its parents' distribution support*

$\text{supp}(\text{Pa}(z)|\mathbf{d})$ *is contained by* $\text{supp}(\text{Pa}(z)|\tilde{\mathbf{d}})$ *for some combination* $\tilde{\mathbf{d}} \in \Omega_{\text{supp}}$ *on the support, i.e.,* $\text{supp}(\text{Pa}(z)|\mathbf{d}) \subseteq \text{supp}(\text{Pa}(z)|\tilde{\mathbf{d}})$.

Proof. By definition, the concept combination $\mathbf{d}$ is composable (i.e., $\mathbf{d} \in \Omega_{\text{comp}}$) when the two alternative model specifications $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}$ agree on this specific $\mathbf{d}$, i.e., $\hat{g}_z = g_z$ for any $z \in \mathbf{z}$ over its inputs' support $\mathcal{S}_z(\mathbf{d}) := \text{supp}(\text{Pa}(z)|\mathbf{d}) \times \text{supp}(\boldsymbol{\epsilon}_z)$. We note that each exogenous variable ϵ_z is independent of $\text{Pa}(z)$ and its distribution remains invariant to the discrete variable $\mathbf{d}$. We denote this relation as $\boldsymbol{\theta}|_{\mathbf{d}} = \hat{\boldsymbol{\theta}}|_{\mathbf{d}}$.

To derive this relation, we first show that under the assumption of the hierarchical data-generating process equation 1, the specific model $\boldsymbol{\theta} := \big(p(\mathbf{z}_1, \mathbf{d}), \{g_v\}_{v \in \mathcal{V} \setminus (\mathbf{z}_1, \mathbf{d})}\big)$'s behavior on the discrete concept space Ω_{comp} is fully determined by its behavior on the support Ω_{supp}. That is, if two specifications $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}$ follow the hierarchical model assumption equation 1 and their behavior match over the support Ω_{supp}, this agreement would extend to Ω_{comp}: $\forall \tilde{\mathbf{d}} \in \Omega_{\text{supp}}, \boldsymbol{\theta}|_{\tilde{\mathbf{d}}} = \hat{\boldsymbol{\theta}}|_{\tilde{\mathbf{d}}} \implies \forall \mathbf{d} \in \Omega_{\text{comp}}, \boldsymbol{\theta}|_{\mathbf{d}} = \hat{\boldsymbol{\theta}}|_{\mathbf{d}}$.

To this end, we assess the elementary generating function $z := g_z(\text{Pa}(z), \epsilon_z)$ for every $z \in \mathbf{z}$ present in the hierarchical model. Although latent variables $\{\mathbf{z}_l\}_{l \in [L+1]}$ form a Markov chain, the first module $p(\mathbf{z}_1|\mathbf{d})$ may yield distinct supports $\text{supp}(\mathbf{z}_1|\mathbf{d})$ across various values of $\mathbf{d}$ (e.g., $d = 0$ for absence of the concept). Consequently, the matching of two models $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}$ is only partially supported and depends on the specific value of $\mathbf{d}$. We characterize a potentially larger composable space Ω_{comp} given their matching over the training support Ω_{supp}. Under the theorem condition, we have $\text{supp}(\text{Pa}(z)|\mathbf{d})$ at the specific $\mathbf{d}$ is fully contained by $\text{supp}(\text{Pa}(z)|\tilde{\mathbf{d}}(\mathbf{d}))$ at some $\tilde{\mathbf{d}}(\mathbf{d}) \in \Omega_{\text{supp}}$ dependent on $\mathbf{d}$, i.e.,

$$\text{supp}(\text{Pa}(z)|\mathbf{d}) \subseteq \text{supp}(\text{Pa}(z)|\tilde{\mathbf{d}}(\mathbf{d})). \tag{7}$$

As the two models g_z and $\hat{g}_z$ match over the discrete support Ω_{supp}, this equality relation in equation 7 implies that this equality extends to $\tilde{\mathbf{d}}(\mathbf{d})$:

$$g_z = \hat{g}_z, \forall (\text{Pa}(z), \boldsymbol{\epsilon}_z) \in \mathcal{S}_z(\tilde{\mathbf{d}}(\mathbf{d})). \tag{8}$$

As the relation in equation 8 holds true for all modules of $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}$, the equality extends to the entire hierarchical model, i.e., $\boldsymbol{\theta}|_{\mathbf{d}} = \hat{\boldsymbol{\theta}}_{\mathbf{d}}$ for $\mathbf{d} \in \Omega_{\text{comp}}$, which concludes our proof.

□

B.2. Proof for Theorem 3.4

Condition 3.3 (Identification Conditions)**.**

i [Invertibility]: There exists a smooth and invertible map $g_l : (\mathbf{z}_l, \boldsymbol{\epsilon}_l) \mapsto \mathbf{x}$ for $l \in [0, L]$.

ii [Smooth Density]: The probability density function $p(\mathbf{z}_{l+1}|\mathbf{z}_l)$ is smooth.

iii [Conditional Independence]: Components in $\mathbf{z}_{l+1}$ are independent given $\mathbf{z}_l$: $p(\mathbf{z}_{l+1}|\mathbf{z}_l) = \prod_n p(z_{l+1,n}|\mathbf{z}_l)$.

iv [Sufficient Variability]: For each value of $\mathbf{z}_{l+1}$, there exist $2n(\mathbf{z}_{l+1}) + 1$ values of $\mathbf{z}_l$, i.e., $\mathbf{z}_l^{(n)}$ with $n = 0, 1, \ldots, 2n(\mathbf{z}_{l+1}) + 1$, such that the $2n(\mathbf{z}_{l+1})$ vectors $\mathbf{w}(\mathbf{z}_{l+1}, \mathbf{z}_l^{(n)}) - \mathbf{w}(\mathbf{z}_{l+1}, \mathbf{z}_l^{(0)})$ are linearly independent, where vector $\mathbf{w}(\mathbf{z}_{l+1}, \mathbf{z}_l)$ is defined as follows:

$$\mathbf{w}(\mathbf{z}_{l+1}, \mathbf{z}_l) = \Big(\frac{\partial \log p\left(\mathbf{z}_{l+1}|\mathbf{z}_l\right)}{\partial z_{l+1,1}}, \ldots, \frac{\partial \log p\left(\mathbf{z}_{l+1}|\mathbf{z}_l\right)}{\partial z_{l+1,n(\mathbf{z}_{l+1})}}, \quad \frac{\partial^2 \log p\left(\mathbf{z}_{l+1}|\mathbf{z}_l\right)}{(\partial z_{l+1,1})^2}, \ldots, \frac{\partial^2 \log p\left(\mathbf{z}_{l+1}|\mathbf{z}_l\right)}{\partial (z_{l+1,n(\mathbf{z}_{l+1})})^2} \Big). \tag{2}$$

Theorem B.1 (Causal Module Identification)**.** *We assume the data-generating process equation 1. Under Condition 3.3, we attain component-wise identifiability of $\mathbf{z}_l$ and the graphical structures $\mathcal{G}$ up to the index permutation at each level l.*

Proof. We introduce Lemma B.1 from Kong et al. [35], which identifies a trivial hierarchical model with only one latent level, i.e., $L = 1$.

Lemma B.1 (Single-level Identification [35])**.** *We assume the following data-generating process equation 1:*

$$\mathbf{z} \sim p(\mathbf{z}|\mathbf{u}), \quad \boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon}), \quad \mathbf{x} := g(\mathbf{z}, \boldsymbol{\epsilon}), \tag{9}$$

where $\boldsymbol{\epsilon}$ refers to the exogenous variable independent of $\mathbf{z}$ and g stands for the generating function. Under Condition 3.3 with $L = 1$ and $\mathbf{z}_0 = \mathbf{u}$, we attain component-wise identifiability of $\mathbf{z}_1$.

In the general hierarchical case, we view the observed discrete variable $\mathbf{d}$ as the top-level variable $\mathbf{u}$ in equation 9 as the starting point. Lemma B.1 implies the component-wise identifiability of $\mathbf{z}_1$. We then iteratively apply Lemma B.1 to identify level $\mathbf{z}_{l+1}$ sequentially from top to bottom equation 1 by viewing the previously identified level $\mathbf{z}_l$ as the conditioning variable $\mathbf{u}$ in equation 9. This reasoning gives the component-wise identifiability results for the entire hierarchical model.

Since all the latent variables $\{z_i\}_{i=1}^{n(\mathbf{z})}$, we can view them as the observed variables. The identifiability of the graphical structure $\mathcal{G}$ follows from classic causal discovery methods (i.e., PC algorithm [58]).

□

C. Additional Details for Experiments

C.1. Setup Details

We train the model with a batch size of 800 and a learning rate of $5e-5$. To inject multiple text conditions, we replicate the key and value linear layers in cross-attention, inspired by IP-Adapter [76]. During testing, we prompt QWEN2.5 [74] with the instruction "given a prompt X, segment it into three non-overlap descriptions (i.e., any two descriptions are not describing the same object), rewrite each subcaption to avoid interactions across each subcaption." For the experiments in Table 3, we employ the LayoutSAM dataset [79] and finetune SANA-1.5 [70] with a batch size of 576 for 20000 steps at a learning rate of $5e-5$. We choose λ from $\{0.1, 1\}$. The performance of **HierDiff** are over three random seeds, with a std of 0.1. **Complexity.** Our method introduces moderate computational overhead. During training, it requires a (pre-computable) LLM forward pass to generate low-level prompts and computes the $\mathcal{L}_{\text{n}}$ sparsity loss, a simple DICE calculation on attention maps. During inference, it requires one LLM pass and M cross-attention computations (where $M = 3$ in our experiments) instead of one at each step. **Training:** fine-tuning requires $\sim$24h on 8×L40 for SD1.5 (UNet backbone), and $\sim$72h on 8×H100 for the DiT-based SANA-1.5. **Inference:** our method runs at 3.94s/4imgs compared to 2.33s/4imgs for SD1.5, and 7.13s/img compared to 5.73s/img for SANA-1.5. This represents a manageable trade-off for the significant improvement in compositional control.

C.2. Language Model Usage

We follow established practices Feng et al. [17], Lian et al. [41], Wu et al. [69], Yang et al. [75] to instruct QWENv2.5 [74] with a fixed instruction. For example, QWENv2.5 rewrites "a peacock is eating ice cream while..." into "A peacock is in the act of eating", ..., "a serving of ice cream is being visibly diminished". In our evaluation, we've found that QWEN2.5 performs decently for most examples, and more advanced models (Gemini 2.5 Flash, Claude 4) are superior on rare, challenging examples involving dense interactions of multiple concepts (e.g., detailed description of multiple mutually overlapping clothing items on a person). To quantify the performance of QWEN2.5, we instruct Claude 4 to evaluate the presence of high-level concepts in captions processed by QWEN2.5 and observe a 96% success rate over 100 DPG evaluation prompts. We believe that the advancement of language models could further improve the performance.

C.3. Additional Samples for Figure 5

Figure 8 and Figure 9 display generated examples from **HierDiff** and baselines, with full text prompts.

C.4. Additional Samples for Figure 6

Figure 10 displays more examples for the ablation experiments in Figure 6.

C.5. More Empirical Understanding

While implicit models can be highly expressive, they can struggle with compositional generalization as many solutions might fit the training data but not generalize beyond. Our work introduces a theoretically motivated sparsity constraint (Eq. 6) to select more generalizable solutions. We include fine-grained qualitative analysis in Fig. 11. In Fig. 11(a), our model attends to "cat" (L1) and "sunglasses" (L2) separately, and the baseline attends to all regions and omits "sunglasses". Similarly, in Fig. 11(b), our model, with sparse constraints focusing attention (L1 on "bear" and L2 on "cat"), renders both; the baseline's simultaneous generation misses "cat". The analysis also highlights cases challenging to our model, such as the difficulty in decomposing words and printing the resultant letters correctly (L1 at 901 covers all letters simultaneously) in Fig. 11(c). While extreme sparsity can affect performance in dense interaction scenes (e.g., missing "herb" in Fig. 11(d)), the model's superior performance over the baseline on examples here and all benchmarks confirms its robustness for these scenarios.

D. Additional Analysis

D.1. Decomposer Ablation

Our method uses a language model to decompose the global prompt into $M=3$ local descriptions (Section 4). To assess whether this LLM dependency is critical, we replace the LLM decomposer with a lightweight heuristic that simply splits the prompt into three equal-length sentence chunks. As shown in Table 4, the heuristic decomposer achieves nearly identical performance to the full LLM-based pipeline on DPG-Bench, and both substantially outperform the recent LLM-based baseline SILMM [50]. This demonstrates that our method's gains stem from the hierarchical injection and sparsity mechanisms, not from LLM prompt rewriting.

Table 4. **Decomposer ablation on DPG-Bench.** Replacing the LLM decomposer with a heuristic yields nearly identical performance.

Configuration	DPG-Bench
HierDiff (heuristic decomposer; no LLM)	79.21
HierDiff (LLM decomposer; full)	**79.28**
SILMM [50] (LLM-based baseline)	77.45

D.2. GenEval Evaluation

To further validate compositional generation beyond DPG-Bench, we evaluate on GenEval [20], a benchmark specifically designed to assess compositional text-to-image alignment through object-focused evaluation. Table 5 shows that **HierDiff** consistently outperforms baselines, confirming our method's advantage on an independent compositional benchmark.

Table 5. **GenEval results (overall accuracy).**

	HierDiff	ELLA	SD1.5
GenEval (Overall)	**0.53**	0.45	0.43

D.3. Disentangling Contributions

To isolate the sources of improvement, we disentangle the contribution of LLM prompt rewriting from our core architectural innovations (hierarchical injection and sparsity regularization). Since ELLA already uses FLAN-T5-XL as its text encoder, the encoder alone cannot explain our gains. Table 6 shows that giving ELLA the same LLM-rewritten prompts adds $+1.90$ points, but our hierarchical injection and sparsity mechanisms yield an *additional* $+2.47$ points. These contributions are additive, confirming that the novelty lies in our framework rather than prompt rewriting.

Table 6. **Disentangling contributions on DPG-Bench.** Prompt rewriting and our core mechanism contribute additively.

Configuration	DPG-Bench	Δ
ELLA (FLAN-T5-XL baseline)	74.91	—
+ LLM prompt rewrite	76.81	+1.90
+ **HierDiff** (hier. inj. + sparsity)	**79.28**	+2.47

D.4. Hyperparameter Sensitivity

We conduct a sensitivity analysis of the sparsity regularization weight λ using the DiT-based model (same setting as Table 3). As shown in Table 7, performance varies by less than 1 point across a wide range of λ values, indicating that **HierDiff** is robust to this hyperparameter. We fix $M=3$ throughout all experiments.

D.5. SAE Verification of Timestep–Hierarchy Correspondence

To empirically verify the connection between diffusion timesteps and the concept hierarchy (Section 4), we train sparse autoencoders (SAEs) on diffusion model representations at early ($t = 15$), mid ($t = 9$), and late ($t = 0$) diffusion steps and

Table 7. λ **sensitivity** on DPG-Bench (DiT-based model).

λ	0.0	0.1	1.0	10.0
DPG-Bench	83.97	**84.91**	84.81	84.15

intervene on the discovered features. As shown in Figure 12, the intervened features exhibit a clear coarse-to-fine hierarchy: early steps encode global concepts (e.g., "full face"), mid steps encode parts (e.g., "nose", "hair"), and late steps encode localized details (e.g., "eyes", "mouth"). This hierarchy is supported by two lines of evidence: (1) the spatial coverage of features is hierarchical (left panel: early features span broad regions, late features are localized), and (2) intervening on a feature changes only its corresponding spatial region (right panel), confirming the hierarchy-level interpretation. This timestep–hierarchy connection is also consistent with findings in prior work [37].

Prompt	SD1.5 [55]	ELLA [24]	Ours
A sleek, black rectangular keyboard lies comfortably on the luxurious beige carpet of a quiet home office, bathed in the gentle sunlight of early afternoon. The keys of the keyboard show signs of frequent use, and it's positioned diagonally across the plush carpet, which is textured with subtle patterns. Nearby, a rolling office chair with a high back and adjustable armrests sits invitingly, hinting at a quick break taken by its usual occupant.			
In the fading light of late afternoon, a scene unfolds in the autumn park, where a pair of worn brown boots stands firm upon a bed of fallen orange leaves. Attached to these boots are two vibrant blue balloons, gently swaying in the cool breeze. The balloons cast soft shadows on the ground, nestled among the trees with their leaves transitioning to auburn hues. Nearby, a wooden bench sits empty, inviting passersby to witness the quiet juxtaposition of the still footwear and the dancing balloons.			
A surreal composite image showcasing the iconic Sydney Opera House with its distinctive white sail-like structures, positioned improbably beside the towering Eiffel Tower, its iron lattice work silhouetted against the night. The backdrop is a vibrant blue sky, pulsating with dynamic energy, where yellow stars burst forth in a dazzling display, and swirls of deeper blue spiral outward. The scene is bathed in an ethereal light that highlights the contrasting textures of the smooth, shell-like tiles of the Opera House and the intricate metalwork of the Eiffel Tower.			
An impressionistic painting depicts a vibrant blue cow standing serenely in a field of delicate white flowers. Adjacent to the cow, there is a robust tree with a canopy of red leaves and branches laden with yellow fruit. The brushstrokes suggest a gentle breeze moving through the scene, and the cow's shadow is cast softly on the green grass beneath it.			
a pyramid-shaped tablet made of a smooth, matte grey stone stands in the foreground, its sharp edges contrasting with the wild, verdant foliage of the surrounding jungle. nearby, a crescent-shaped swing hangs from a sturdy tree branch, crafted from a polished golden wood that glimmers slightly under the dappled sunlight filtering through the dense canopy above. the swing's smooth surface and gentle curve invite a sense of calm amidst the lush greenery.			
In a spacious loft with high ceilings and exposed brick walls, the morning light filters through large windows, casting a soft glow on a pair of trendy, high-top sneakers. These sneakers, made of rugged leather with bold laces, contrast sharply with the ornate, metallic vintage coffee machine standing next to them. The coffee machine, with its intricate details and polished finish, reflects the light beautifully, setting a striking juxtaposition against the practical, street-style footwear on the polished concrete floor.			

Figure 8. **More text-to-image generation results.**

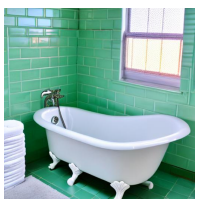
Prompt	SD1.5 [55]	ELLA [24]	Ours
An elegant pair of glasses with a unique, gold hexagonal frame laying on a smooth, dark wooden surface. The thin metal glints in the ambient light, highlighting the craftsmanship of the frame. The clear lenses reflect a faint image of the room's ceiling lights. To the side of the glasses, a leather-bound book is partially open, its pages untouched.			
Two multicolored butterflies with delicate, veined wings gently balance atop a vibrant, orange tangerine in a bustling garden. The tangerine, with its glossy, dimpled texture, is situated on a wooden table, contrasting with the greenery of the surrounding foliage and flowers. The butterflies, appearing nearly small in comparison, add a touch of grace to the scene, complementing the natural colors of the verdant backdrop.			
Two sleek blue showerheads, mounted against a backdrop of white ceramic tiles, release a steady stream of water. The water cascades down onto ORANGEa vivid, crisp green pearthat is centrally positioned directly beneath them. The pear's smooth and shiny surface gleams as the water droplets rhythmically bounce off, creating a tranquil, almost rhythmic sound in the otherwise silent bathroom.			
In a modern kitchen, a square, chrome toaster with a sleek finish sits prominently on the marble countertop, its size dwarfing the nearby red vintage rotary telephone, which is placed quaintly on a wooden dining table. The telephone's vibrant red hue contrasts with the neutral tones of the kitchen, and its cord coils gracefully beside it. The polished surfaces of both the toaster and the telephone catch the ambient light, adding a subtle shine to their respective textures.			
Two slender bamboo-colored chopsticks lie diagonally atop a smooth, round wooden cutting board with a rich grain pattern. The chopsticks, tapered to fine points, create a striking contrast against the cutting board's more robust and circular form. Around the board, there are flecks of freshly chopped green herbs and a small pile of julienned carrots, adding a touch of color to the scene.			
A cozy bathroom features a pristine, white claw-foot bathtub on a backdrop of pastel green tiles. Adjacent to the tub, a tower of soft, white toilet paper is neatly stacked, glimmering gently in the diffuse glow of the afternoon sunlight streaming through a frosted window. The gentle curvature of the tub contrasts with the straight lines of the stack, creating a harmonious balance of shapes within the intimate space.			

Figure 9. **More text-to-image generation results.**

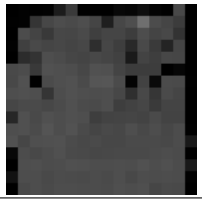
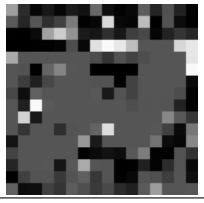
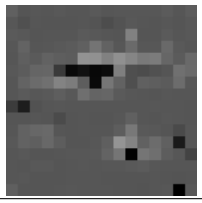
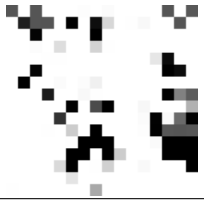
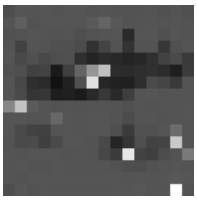
Prompt	w/o TD	w/o SR	**HierDiff**
A vibrant pink pig trots through a snowy landscape, a bright blue backpack strapped securely to its back. The pig's thick coat contrasts with the soft white blanket of snow that covers the ground around it. As it moves, the blue backpack stands out against the pig's colorful hide and the winter scene, creating a striking visual amidst the serene, frost-covered backdrop.			
			
			
An outsized dolphin with a sleek, gray body glides through the blue waters, while a small, fluffy chicken with speckled brown and white feathers stands on the nearby sandy shore, appearing diminutive in comparison. The dolphin's fins cut through the water, creating gentle ripples, while the chicken pecks at the ground, seemingly oblivious to the vast size difference. The stark contrast between the dolphin's smooth, aquatic grace and the chicken's terrestrial, feathered form is highlighted by their proximity to one another.			
			
			

Figure 10. **More ablation studies.** Without time-dependence (TD), the model fails to understand the relationship among the objects in the prompt. Without sparsity regularization (SR), the influence of each prompt could be large, e.g., the attention map of local prompt 1 covers the pineapple and beers. Combining the two proposed designs, **HierDiff** generates images that accurately follow the complex text prompt.

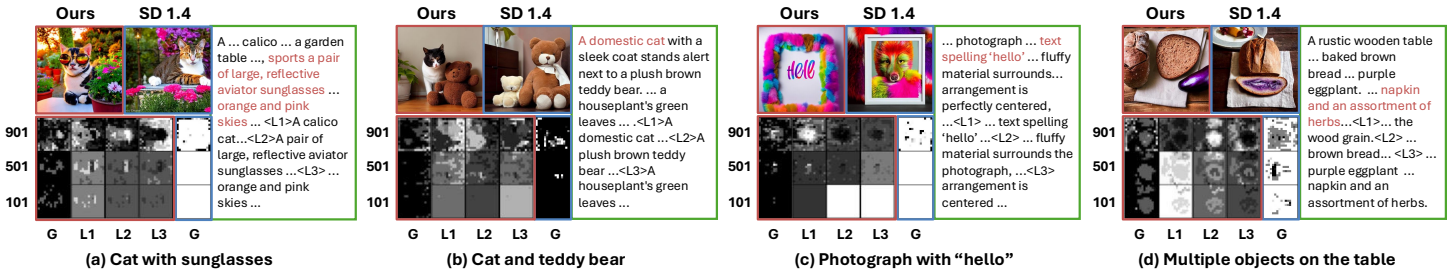


Figure 11. **Fine-grained comparison between our method and Stable Diffusion 1.4.** G and L_i indicate full caption and split captions (for our method), and indices denote diffusion steps (901 is closer to noise). White indicates high attention scores.

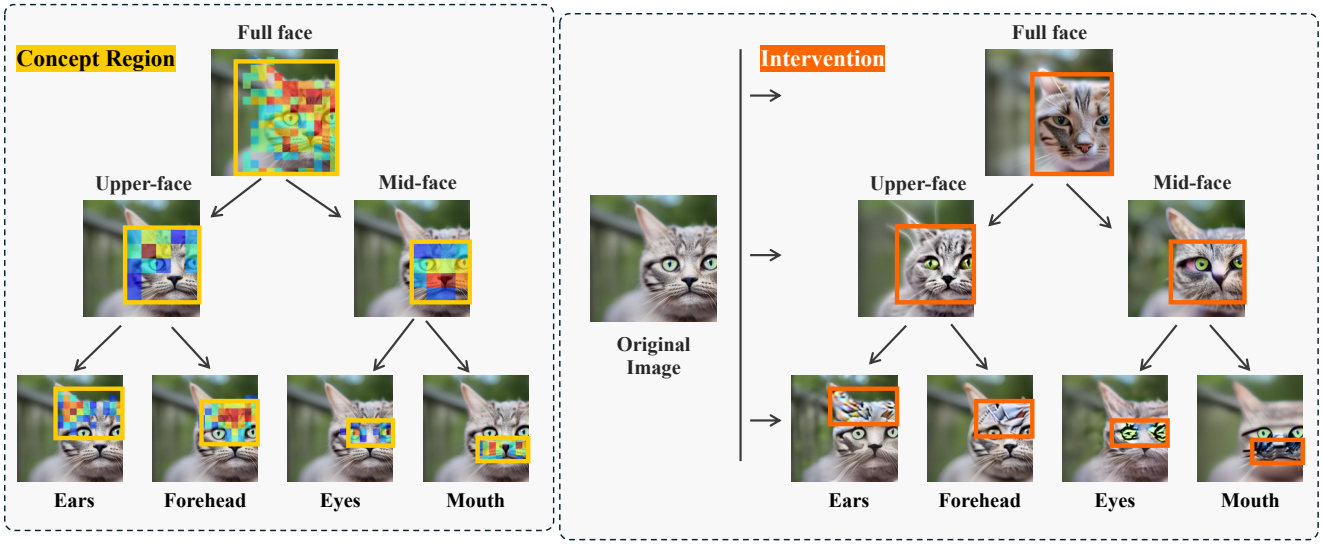


Figure 12. **SAE evidence for coarse-to-fine concepts across diffusion steps.** Left: hierarchical concept graph discovered from sparse features at different timesteps. Right: intervening on a sparse feature (e.g., "eyes") changes only its corresponding spatial region, supporting the hierarchy-level interpretation.