# Synthia: Novel Concept Design with Affordance Composition

**Hyeonjeong Ha**[1*], **Xiaomeng Jin**[1*], **Jeonghwan Kim**[1],
**Jiateng Liu**[1], **Zhenhailong Wang**[1], **Khanh Duy Nguyen**[1], **Ansel Blume**[1],
**Nanyun Peng**[2], **Kai-Wei Chang**[2], **Heng Ji**[1]
[1]University of Illinois Urbana-Champaign
[2]University of California Los Angeles
{hh38, xjin17, hengji}@cs.illinois.edu

## Abstract

Text-to-image (T2I) models enable rapid concept design, making them widely used in AI-driven design. While recent studies focus on generating semantic and stylistic variations of given design concepts, *functional coherence*—the integration of multiple affordances into a single coherent concept—remains largely overlooked. In this paper, we introduce **Synthia**, a framework for generating visually novel and functionally coherent designs based on desired affordances. Our approach leverages a hierarchical concept ontology that decomposes concepts into parts and affordances, serving as a crucial building block for functionally coherent design. We also develop a curriculum learning scheme based on our ontology that contrastively fine-tunes T2I models to progressively learn affordance composition while maintaining visual novelty. To elaborate, we (i) gradually increase affordance distance, guiding models from basic concept-affordance association to complex affordance compositions that integrate parts of distinct affordances into a single, coherent form, and (ii) enforce visual novelty by employing contrastive objectives to push learned representations away from existing concepts. Experimental results show that Synthia outperforms state-of-the-art T2I models, demonstrating absolute gains of 25.1% and 14.7% for novelty and functional coherence in human evaluation, respectively. Code is available at https://github.com/HyeonjeongHa/SYNTHIA.

## 1 Introduction

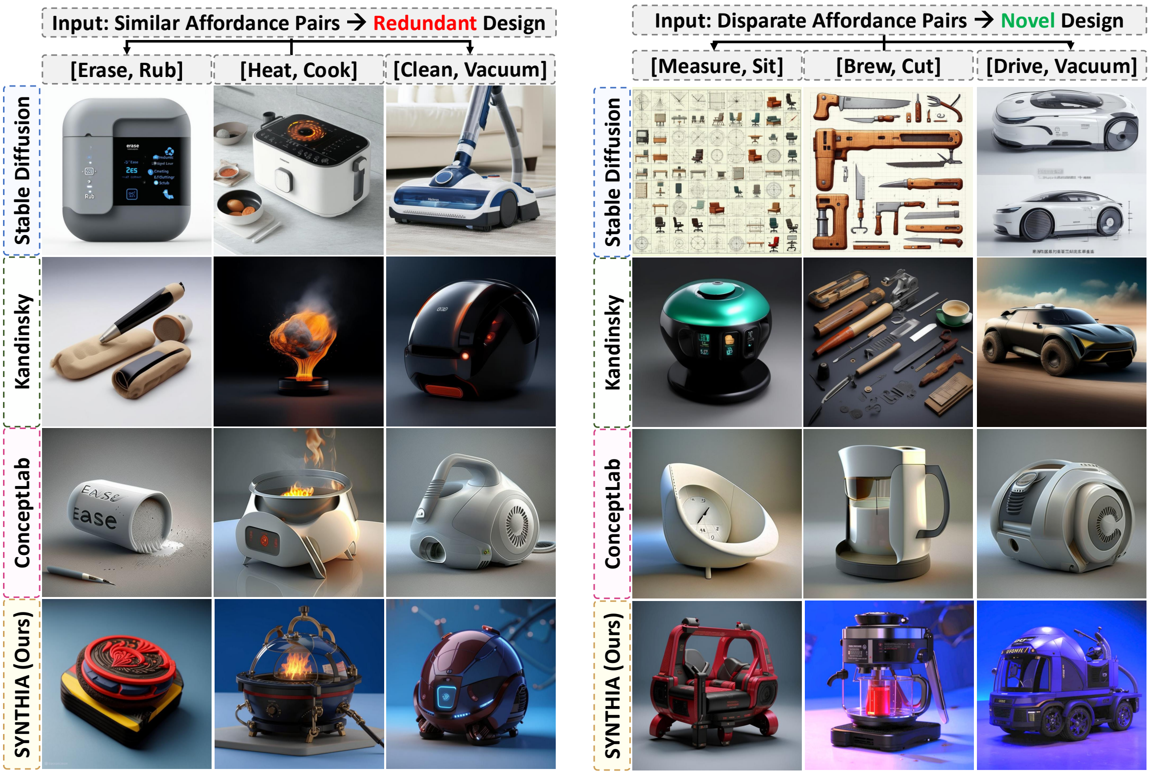
Imagine a coffee machine with wheels that brews your morning coffee and delivers it directly to your bed. This example illustrates a novel concept that is atypical and dissimilar to everyday objects we regularly encounter. Such *novel concept design* is defined as the synthesis of concepts that are visually unique and functionally coherent through the effective fusion of disparate concepts (e.g., `coffee machine`, `trolley`) based on the user's desired affordances (e.g., `brew`, `deliver`). This process mirrors how humans blend ideas across cognitive domains to generate creative innovations (Fauconnier and Turner, 2002; Han et al., 2018).

Existing studies on conceptual design using T2I models have enabled rapid ideation of novel visual concepts (Cai et al., 2023; Ma et al., 2023; Wang et al., 2024; Lin et al., 2025) by identifying user challenges such as interpreting abstract concepts in language to help visualize a novel design concept (Lin et al., 2025), or using large language models (LLMs) to bootstrap initial ideation in texts (Cai et al., 2023; Zhu and Luo, 2023). However, they often naively feed LLM-generated textual prompts into T2I models, relying on simple key phrases or semantic variations of concept description (Cai et al., 2023; Wang et al., 2024). While existing works show that T2I models can generate images that seem to correctly reflect complex human-formulated textual descriptions (e.g., *"beautiful rendering of neon lights in futuristic cyberpunk city"*), they do not focus on whether a model can synthesize a novel concept when given a set of affordances as input, while ensuring these affordances are preserved.

An important aspect lacking in existing approaches to concept synthesis is their focus on pixel-based control, overlooking the structural and functional roles embedded in design. Many real-world concepts are naturally "decomposable" into parts, where each part signals a specific functionality. To address this, we propose **Synthia**, a framework for Concept **Synth**es**i**s with **A**ffordance composition that generates functionally coherent and visually novel concepts given a set of desired affordances. Unlike prior works relying on complex descriptive text to generate stylistic variations or

[*] Equal contribution.

*Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20939–20958
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics



(a) Generated concepts with similar affordances.

(b) Generated concepts with distant affordances.

Figure 1: **Effect of Affordance Sampling on Novel Concept Generation.** Our affordance sampling strategy selects disparate affordance pairs within our ontology, promoting novel functional coherence rather than redundant combinations. Baseline models tend to generate existing concepts for close affordances (Fig. 1a) but struggle with distant pairs, often introducing multiple objects or omitting functions (Fig. 1b). In contrast, our models consistently generate functionally coherent novel concepts, achieving higher novelty scores for distant affordance pairs.

aesthetic features (Richardson et al., 2024; Vinker et al., 2023), SYNTHIA leverages affordances–defined as “the functionality offered by an object or its parts”—as a structural guide for novel concept synthesis. By aligning textual descriptions with affordances as control signals, our models implicitly learn to “decompose and reassemble” functional parts, ensuring that, for instance, a hybrid of a coffee machine and a trolley not only appears novel but also retains its brewing and mobility functions, achieving *functional coherence*.

To facilitate structured affordance composition, we construct a hierarchical concept ontology that decomposes visual concepts (e.g., Sofa) into their constituent parts (e.g., leg, cushion) and associated affordances (e.g., support, rest). It provides a structured representation of concept-affordance associations, serving as the foundation for generating functionally meaningful designs. Inspired by the theory of combinational creativity in humans (Han et al., 2018), which suggests novel concepts emerge from disparate ideas, we propose an affordance sampling mechanism that strategically selects affordances associated with sufficiently different concepts using our novel similarity-based metric (§3.1). This ensures that generated designs integrate novel functionalities, avoiding trivial combinations, whereas random sampling yields similar affordances (e.g., cook, heat) that result in redundant outputs (Fig. 1a).

We also introduce a new curriculum learning scheme that fine-tunes T2I models to progressively learn affordance composition while maintaining visual novelty. Our curriculum gradually increases the affordance distance, allowing models to first learn basic concept-affordance associations from close affordance pairs before tackling complex affordance compositions that integrate multiple affordances into a single, coherent form. To further ensure novelty, we employ contrastive objectives to push learned representations away from existing concepts in our ontology. This addresses a critical limitation of existing T2I models, which struggle to generate coherent multi-functional concepts (Fig. 1b). Without structured affordance composition, models tend to default to familiar objects—e.g., when prompted with `drive` and `vacuum` affordances, Stable Diffusion models simply generate a car with missing `vacuum` functions (Fig. 1b), rather than blending both affordances into an integrated design. Importantly, unlike existing AI-driven design frameworks that rely on detailed LLM-generated descriptions, SYNTHIA enables direct affordance-based prompting, e.g., "*a new design that has functions of {desired affordances}.*". Our model implicitly learns concept-affordance associations, producing novel, structured designs without redundant textual prompting.

To evaluate our framework, we uniformly sample 500 unseen affordance pairs from our ontology and assess generated concepts using automatic and human evaluation. We design evaluation metrics (§4.2) that measure faithfulness, novelty, practicality, and coherence. Experiments show that SYNTHIA significantly outperforms baselines, creating designs that are visually novel and functionally coherent, with consistently higher scores across all metrics. Our contributions are as follows:

- We introduce a hierarchical concept ontology that encodes concept-affordance associations, serving as crucial building blocks for novel concept synthesis with functional coherence.
- We propose an affordance sampling strategy that guides disparate affordance selection, avoiding redundant functionalities while ensuring coherent concept synthesis.
- We develop a curriculum-based optimization for affordance composition that fine-tunes T2I models, enabling T2I models to fuse multiple affordances into a single coherent concept.

## 2 Related Work

**Text-to-Image Models.** The advancement of text-to-image (T2I) models has enabled high-quality image synthesis from textual descriptions (Sohn et al., 2023; Xue et al., 2024; Shi et al., 2024; Chen et al., 2024). Especially, the invention of diffusion-based models, such as DALL-E (Ramesh et al., 2021a) and Stable Diffusion (Rombach et al., 2022), significantly increases the performance of the T2I generation by utilizing a transformer-based architecture, where the image embeddings and text encodings are aligned in the shared representation space. For instance, Bao et al. (2024) propose a compositional fine-tuning method for T2I Diffusion Models that focuses on two novel objectives and performs fine-tuning on critical parameters. However, these models still struggle to understand practical functionalities and integrate multiple components into coherent novel concepts. This highlights the need for a new framework to enhance the compositional reasoning ability of T2I models, which our work aims to address.

**Novel Concept Generation.** The great power of T2I models provides a potential boost to content creation (Ko et al., 2023; Rangwani et al., 2024; Sankar and Sen, 2024; Rahman et al., 2024; Tang et al., 2024; Zhou et al., 2025). Novel concept generation aims to produce visual outputs that extend the existing concepts by specifying the requirements as input to the T2I models. For instance, Concept Weaver (Kwon et al., 2024) first generates a template image based on a text prompt, then refines it using a concept fusion strategy. ConceptLab (Richardson et al., 2024) utilizes Diffusion Prior models and formulates the generation process as an optimization process over the output space of the diffusion prior. Yet, they focus on concept-level generation and ignore the relationships between concepts and their parts. By prioritizing aesthetics, they limit real-world practicality. Our work bridges this gap by designing an affordance-driven novel concept synthesis framework, integrating desired functions to output novel but practical concepts.

## 3 SYNTHIA: Novel Concept Design with Affordance Composition

Our ultimate goal is to utilize T2I models in designing novel concepts that are both visually novel and functionally coherent. Specifically, we take the desired affordances as text inputs, the T2I models should generate an image that depicts the novel concept design. To achieve this, we (1) construct a training recipe that explicitly embeds hierarchical relations on visual concepts, parts, and corresponding affordances, and (2) fine-tune T2I models with curriculum-based optimization (Fig. 2).

### 3.1 Affordance Composition Curriculum

The primary challenge in the novel concept generation of existing T2I models is the lack of structured functional grounding. These models often struggle to design visually novel yet functionally coherent concepts while maintaining intended functionalities. For example, when combining affordances like `Brew` and `Cut`, they may prioritize aesthetics over functionality, omitting parts or objects

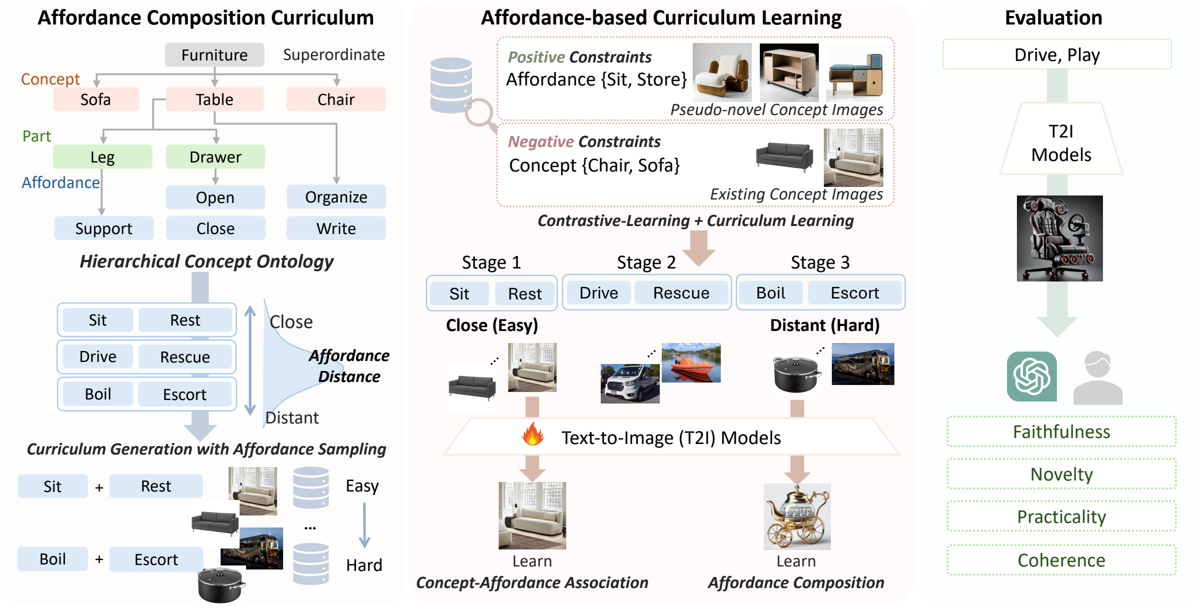


Figure 2: **SYNTHIA: Novel Concept Design with Affordance Composition.** SYNTHIA comprises three stages: (1) Affordance composition curriculum construction, (2) Affordance-based curriculum learning, and (3) Evaluation. In the first stage, we build a training curriculum through sampling affordance pairs from our ontology by gradually increasing the affordance distances. Using our curriculum, we fine-tune T2I models, where they first learn concept-affordance associations from easy data, then integrate multiple affordances into a single functional form from hard data. We employ a contrastive objective with positive (affordances), negative (concepts) constraints, and corresponding images, enforcing visual novelty different from existing concepts. Finally, we evaluate models through automatic evaluation and human evaluation with four metrics: faithfulness, and novelty, practicality, coherence.

relevant to `Brew` (Fig. 1b). To address this, we construct a structured training recipe in two key steps: (1) building a hierarchical concept ontology, and (2) designing an affordance sampling strategy for curriculum-based training. This improves the model's composition ability by learning the connection between concepts and affordances.

**Hierarchical Concept Ontology.** To provide a structured basis for novel concept synthesis with functional coherence, we define a hierarchical concept ontology that decomposes visual concepts into constituent parts and their affordances, capturing concept-affordance associations (Fig. 6). This ontology allows T2I models to retrieve relevant parts based on affordances, enabling generation to be well-grounded on the functionality of concepts rather than superficial visual feature combinations. Formally, we define the ontology as a four-level hierarchy $\mathcal{O} = (\mathcal{S}, \mathcal{C}, \mathcal{P}, \mathcal{A})$. The superordinate $\mathcal{S}$ denotes the highest-level categories, such as `furniture`, followed by Concept $\mathcal{C}$, which is the set of visual concepts. Each $c \in \mathcal{C}$ belongs to a superordinate category $s \in \mathcal{S}$, e.g., $\mathcal{S}_{\texttt{table}} = \texttt{furniture}$, and decomposes into its parts $\mathcal{P}$ that serve specific functions in an object design, e.g., $\mathcal{P}_{\texttt{table}} = \{\texttt{leg}, \texttt{drawer}\}$. The Affordance $\mathcal{A}$ describes functionalities of concepts and parts. Both a concept $c \in C$ and its part $p \in P$ are linked to affordances set $\mathcal{A}_c = \{a_1, \cdots, a_n\} \in \mathcal{A}$, e.g, $\mathcal{A}_{\texttt{table}} = \{\texttt{write}, \texttt{organize}\}$, and $\mathcal{A}_p = \{p_1, \cdots, p_n\}$, e.g., $\mathcal{A}_{\texttt{leg}} = \{\texttt{support}\}$. Our ontology spans 30 superordinates, 590 concepts, 1172 parts, and 686 affordances, explicitly providing a structured representation of how affordance is associated with fine-grained parts for functionally grounded novel concept synthesis.

**Affordance Sampling.** Given our ontology, we can utilize it to create fine-tuning data to improve the functional coherence of the novel concept generated by the T2I models. A naive approach to obtaining training data would be to exhaustively pair all possible affordances. However, this would yield 235K affordance pairs, which is computationally expensive. Moreover, random combination risks generating redundant concepts (e.g., `Heat` and `Cook` examples in Fig. 1a) or functionally incoherent objects. To achieve sufficiently different affordance pairs that enable novel concept synthesis while still being functionally integrable, we introduce a distance-based affordance sampling strategy that selects meaningful, disparate affordance pairs based on ontology-derived distances.

We define a concept distance $D_C(c_i, c_j)$ between two concepts $c_i, c_j \in \mathcal{C}$ by incorporating functional relatedness at the affordance level and semantic similarity at the concept level. We compute functional relatedness using Jaccard similarity $J(X, Y) = |X \cap Y|/|X \cup Y|$ between their affordance sets while quantifying semantic similarity Sim by measuring embedding similarity using the BERT (Devlin et al., 2019) model as follows:

$$D_{\mathcal{C}}(c_i, c_j) = \alpha * \{J(\mathcal{A}_{c_i}, \mathcal{A}_{c_j}) + J(\mathcal{A}_{P_{c_i}}, \mathcal{A}_{P_{c_j}})\}, \\ + \beta * \text{Sim}(\text{BERT}(c_i), \text{BERT}(c_j)), \tag{1}$$

where $\alpha, \beta$ are adjustable hyperparameters that balance between functional relatedness based on affordances, and semantic relevance of concepts, respectively. Since we prioritize affordance-level similarity over concept-level similarity during training, we set $\alpha = 0.7$ and $\beta = 0.3$. Two semantically similar concepts sharing more affordances have closer distances, such as `sofa` and `chair`, while those that have different affordances and semantic differences, such as `car` and `vacuum cleaner` have more distant distances.

We further obtain the affordance distance $D_{\mathcal{A}}(a_i, a_j)$ between two affordances $a_i, a_j \in \mathcal{A}$ by averaging the pairwise concept distances $D_{\mathcal{C}}(\cdot, \cdot)$ between associated concepts:

$$D_{\mathcal{A}}(a_i, a_j) = \frac{1}{|C_{a_i}| \cdot |C_{a_j}|} \sum_{c_p \in C_{a_i}} \sum_{c_q \in C_{a_j}} D_C(c_p, c_q), \tag{2}$$

where $C_{a_i}$ and $C_{a_j}$ are the sets of concepts associated with affordances $a_i$ and $a_j$, respectively. The resulting $D_{\mathcal{A}}(\cdot, \cdot)$ is distributed from 0.1 to 1.0.

Based on our distance metric, close affordance pairs associated with similar concepts, e.g., {`sit`, `rest`} from {`sofa`, `chair`}, support learning basic affordance-concept associations, which can be easily merged into existing concepts. In contrast, distant affordance pairs derived from sufficiently distant concepts, e.g., {`drive`, `vacuum`} from {`car`, `vacuum cleaner`}, enforce greater functional coherence by requiring meaningful part-affordance integration, which is more complex than a trivial combination.

**Curriculum Construction.** In novel concept generation, existing T2I models struggle with (1) concept-affordance associations and (2) the composition of functionally coherent affordances into a single concept. To address these challenges with limited data, we propose a three-stage curriculum that progressively increases affordance pair distances. In the earliest stage, we utilize close affordance pairs to reinforce fundamental knowledge of the concept-affordance associations. The second stage employs the affordance pairs from the mid-range distances to encourage the model to learn the fine-grained compositional structures while maintaining prior knowledge. In the last stage, we only introduce distant affordance pairs to challenge the model to synthesize novel, functionally coherent concepts by applying the previously learned basics on the fine-grained parts and affordance relations.

We sample 600 affordance pairs uniformly across the full distance spectrum and categorize them into three groups. For training images used as pseudo novel concepts, we generate 10 images per pair using DALL-E (Ramesh et al., 2021b) with GPT-4o (OpenAI, 2024) generated captions that describe different novel designs integrating the specified affordances (Details are described in the Appendix B.1). We then filter images using CLIP similarity scores and manually select the top three. This curriculum-based training enables T2I models to learn basic concept-affordance associations while fusing affordances into coherent and functionally meaningful designs. Thus, the T2I models can successfully produce novel concepts that are visually distinctive and functionally coherent.

### 3.2 Affordance-based Curriculum Learning

The goal of fine-tuning T2I models is to enable them to fuse multiple affordances into a single, functionally coherent concept while ensuring visual novelty. With our curriculum, we propose a curriculum learning strategy to fine-tune the diffusion-based T2I models. From a data-driven perspective, training with affordance pairs and DALL-E-generated pseudo-novel concepts helps the model design novel concepts given specified affordances.

To further enhance visual novelty, we incorporate contrastive learning objectives, ensuring that generated images not only reflect desired affordances but also differ from existing concepts associated with them. Specifically, we define two sets of constraints derived from our ontology to guide the model: (1) *Positive Constraints* specify the target affordances that must be included in the novel concepts, shaping their functional structure;

(2) *Negative Constraints* consist of all existing concepts from our ontology that already have the target affordances in the positive constraints. These act as references to avoid. By adhering to these constraints, the model generates concepts that successfully integrate the specified affordances while maintaining a high degree of novelty.

**Training Objectives.** The training objective of fine-tuning is formulated using a triplet loss, which can balance two components to achieve the desired outcomes. The first component aims to minimize the similarity loss between the generated image and the pseudo-novel image created during curriculum construction, ensuring visual novelty. To reduce the overfitting problem, we also sample multiple pseudo-novel images that describe different concepts. Given the set of affordances $\mathcal{A}_{pos} = \{a_1, \cdots, a_n\}$ in the positive constraints, together with a sampled image $I_i^+$ from the pseudo-novel images from DALL-E, the positive loss is defined as follows:

$$\mathcal{L}_{pos}(\theta_t) = \|I_i^+ - \hat{I}_i\|_2^2 + \mathbb{E}_{\epsilon,t}\left[\|\epsilon - \epsilon_\theta(t)\|_2^2\right], \quad (3)$$

where $\theta_t$ is T2I model parameters, $\hat{I}_i$ denotes the generated image, $\epsilon$ is Gaussian random noise. We employ noise prediction loss, where the model takes the latent embedding of $I_i^+$ as input and predicts the noise as $\epsilon_\theta(t)$, preventing catastrophic forgetting of learned training distribution.

The second component of the triplet loss maximizes the similarity loss between the generated image and a randomly sampled existing concept image $I_i^-$ that contains partial affordances from the positive constraints as follows:

$$\mathcal{L}_{neg}(\theta_t) = \|I_i^- - \hat{I}_i\|_2^2 \quad (4)$$

In this way, the model learns to avoid generating existing concept images and increase its novelty.

Our overall triplet loss is defined as follows:

$$\mathcal{L}(\theta_t) = \mathcal{L}_{pos}(\theta_t) - \gamma * \mathcal{L}_{neg}(\theta_t), \quad (5)$$

where $\gamma$ is an adjustable hyperparameter. By balancing two losses, our framework ensures that the generated images align with the desired affordances while remaining distinct from existing concepts.

### 3.3 Novel Concept Generation during Inference

After fine-tuning the diffusion-based T2I models, our approach requires only the desired affordances as positive constraints during inference time, eliminating the need for manually collecting existing concepts as negative constraints. This efficiency gain stems from incorporating both positive and negative constraints–derived from our hierarchical concept ontology–into the training objective. By embedding these constraints during training, the model learns concept-affordance associations and improves its ability to compose parts associated with desired affordances into a novel design. Therefore, the model can produce novel, structured designs without redundant textual prompting.

## 4 Experiments

### 4.1 Experimental Setup

**Datasets.** To train T2I models with our approach, we construct a dataset from two types of resources (more details in the Appendix B.2): (1) *Existing Concept Images*: For each existing concept in our ontology, we collect 60 images from external platforms including Google Images and iStock. To ensure that images are object-centric and aligned with the concept, we filter out low quality images using the CLIP model (Radford et al., 2021). (2) *Generated Novel Concept Images*: With our affordance sampling, we uniformly sample 600 affordance pairs among 235K possible pairs for fine-tuning. For the test dataset, we select 500 affordance pairs among the ones not used for fine-tuning.

**Baselines Methods.** We compare our proposed method against three baseline methods, which are Stable Diffusion (Esser et al., 2024), Kandinsky (Arkhipkin et al., 2023), and ConceptLab (Richardson et al., 2024). While Stable Diffusion and Kandinsky are general T2I models, ConceptLab optimizes generation over diffusion before creative concept design. For a fair comparison, we fine-tune ConceptLab using the same training data as our method. In contrast, our framework directly fine-tuned the diffusion model, integrating the hierarchical visual ontology to enforce the design of a single, coherent concept that achieves multiple affordances. Details on the baselines can be found in the Appendix B.3.

**Implementation Details.** We adopt Kandinsky3.0 (Vladimir et al., 2024) as the T2I backbone model; it generates images based on a given text prompt, with an optional negative text prompt to refine outputs. During fine-tuning, we incorporate the desired affordances as positive inputs, while

|  | Automatic Evaluation |  |  |  | Human Evaluation |  |  |  |
|---|---|---|---|---|---|---|---|---|
| **Model** | **Faithfulness** | **Novelty** | **Practicality** | **Coherence** | **Faithfulness** | **Novelty** | **Practicality** | **Coherence** |
| Stable Diffusion | 3.77 | 3.74 | 3.34 | 3.29 | 2.96 | 2.44 | 3.02 | 2.75 |
| Kandinsky3 | 3.38 | 4.02 | 2.92 | 3.89 | 2.95 | 2.98 | 3.01 | 3.41 |
| ConceptLab | 3.39 | 4.08 | 2.93 | 3.96 | 2.73 | 3.11 | 2.68 | 3.54 |
| Synthia (Ours) | **3.99** | **4.55** | **3.35** | **4.81** | **3.81** | **3.89** | **3.38** | **4.06** |

Table 1: **Results of the automatic evaluation and human evaluation.** We compare our method with baseline models. Each metric ranges from 0 to 5, where a higher score indicates a better performance.

using the existing concepts from the ontology as negative constraints. During inference, we provide only text prompts with desired affordances, "*a new design that has functions of {desired affordances}.*". Training details are provided in the Appendix B.4

### 4.2 Evaluation Metrics

**Automatic Evaluation.** We adopt the LLM-as-a-Judge (Zheng et al., 2023; Kim et al., 2023) for automatic evaluation, due to its strong alignment with human judgments. We especially utilize a proprietary model, GPT-4o (OpenAI, 2024), capable of processing both visual and textual inputs. Building on this, we design four novel metrics to automatically assess the quality of the generated data:

- **Faithfulness:** This metric evaluates how well the generated object aligns with instructions, focusing on its intended affordances and whether the image effectively conveys the object's purpose.
- **Novelty** assesses the originality and creativity of the generated design, emphasizing uniqueness and unconventional concepts that surprise or intrigue users.
- **Practicality** evaluates the real-world applicability of the design. It examines usability, alignment with human preferences, and feasibility for production.
- **Coherence** evaluates whether the generated image is object-centric, depicting a single clear and functional object without unintended elements. It examines whether multiple affordances are fused into a unified concept rather than shown as separate objects.

For all four metrics, we use absolute scores ranging from 1 to 5, with higher scores indicating better quality. However, since the scores for these metrics may be influenced by subjective interpretation, we also include a relative evaluation. Specifically, we present each generated image with its corresponding DALL-E generated image, and ask the automatic evaluator to compare and determine which is superior or if they are equally strong. This relative comparison ensures a more fair evaluation and reduces potential biases (detailed in the Appendix B.5).

**Human Evaluation.** To assess the quality of the generated concepts beyond automated evaluations, we conduct a human evaluation with 36 non-design expert annotators. Recruited from the university across diverse majors, they are provided with a detailed rubric using the same metrics and a 1-5 scale as automated evaluations. We randomly sample 10 affordance pairs for four models, with each sample independently evaluated. This allows direct comparisons between human and automated scores, capturing nuanced aspects of evaluation quality. Details of the evaluation process are documented in the Appendix B.6 to ensure transparency.

### 4.3 Results and Analysis

**Automatic Evaluation.** We compare Synthia against three existing T2I models using our evaluation metrics. To ensure a fair comparison, we randomly sample 500 test affordance pairs not seen during training. As shown in the absolute evaluation results (Table 1), Stable Diffusion consistently achieves high practicality but lower novelty. This aligns with our observation that it tends to generate existing concepts rather than novel ones (Fig. 1). When it fails to retrieve an existing concept that satisfies all the affordances in the prompt, it often generates multiple disjoint objects without meaningful fusion, leading to lower coherence scores for both concept and affordance levels. In contrast, Kandinsky-3 and ConceptLab exhibit improved novelty and coherence compared to Stable Diffusion, but at the cost of reduced practicality. Their outputs are often more imaginative but less grounded in affordances. Our method, Synthia, outperforms all baselines in faithfulness, novelty,

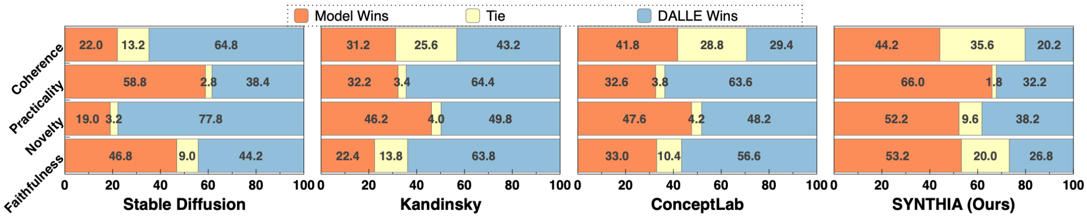


Figure 3: **Results of the relative automatic evaluation.** We compare the quality of concepts generated from our models and baselines with ones generated from our data generation pipeline (§3.1). Numbers indicate the percentage (%) of baseline model wins, ties, and DALL-E model wins.

and coherence, while also maintaining strong performance in practicality. These findings are reinforced by the relative evaluation results (Fig. 3 and 7), where SYNTHIA consistently achieves higher win rates against DALL-E across all metrics compared to baselines. Notably, SYNTHIA even surpasses DALL-E in overall preference, indicating that its generated concepts are perceived as more innovative and functionally coherent. These results demonstrates that fine-tuning with our curriculum strategy enables SYNTHIA to better follow textual instructions, effectively integrate multiple affordances, and generate high-quality novel concepts.

**Human Evaluation.** To assess the consistency of human evaluations, we compute inter-annotator agreement (IAA) between two independent raters, considering ratings to be in agreement if their absolute difference is $\leq 1$. Overall IAA across all images was 67.5%, with a Cohen's Kappa of 22.3%. Among the evaluation criteria, Novelty shows the highest agreement (70.9%), followed by Faithfulness (68.5%), Practicality (66.4%), and Coherence (64.1%) (Fig. 8). Additionally, the agreement between aggregated human ratings and automatic evaluations reached 91.25%, suggesting strong alignment between human judgments and our automated evaluation framework. These results reinforce that human annotators demonstrate a reasonable level of consistency, and that our automatic evaluation metrics reliably reflect human preferences. Human evaluation further confirms that our model consistently generates functionally coherent and visually novel concepts, achieving superior performance across all metrics (Table 1).

## 4.4 Ablation Studies

**The Size of Training Data.** In our experiment, we fine-tune the diffusion model using 600 affordance pairs as training data. To investigate the impact of training data size, we compare model performance across different dataset scales: 200, 400, 600, and 800 affordance pairs, evaluated using our automatic evaluation protocol. As shown in Fig. 4, performance improves as the dataset size increases, reaching its peak at 600 pairs. Notably, our model consistently outperforms all baseline methods across all training sizes, demonstrating the efficacy of our approach.

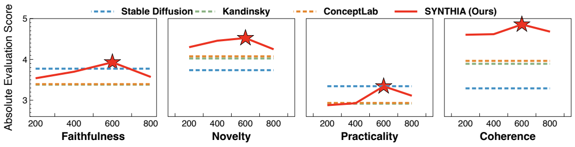


Figure 4: **Ablation with different number of training data.** We show the absolute automatic evaluation results of SYNTHIA trained with different number of data.

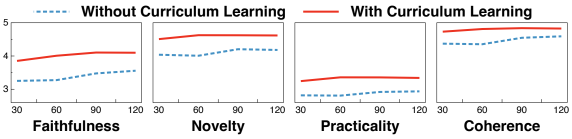


Figure 5: **Effectiveness of curriculum learning.** We show learning curves of SYNTHIA with different training methods. The X-axis represents training steps.

**Effectiveness of Affordance Sampling.** To examine how the distance between affordance pairs affects concept novelty, we sample 100 test pairs with the closest and 100 with the farthest distances. The automatic novelty scores for each group, shown in Table 6, demonstrate that all three baseline methods achieve relatively low novelty scores for closely related affordance pairs, indicating a tendency to replicate existing concepts rather than generating novel designs. In contrast, our method consistently exhibits high novelty across various distances, outperforming all baselines and demonstrating its ability to generate creative and diverse concepts regardless of affordance similarity.

**Effectiveness of Curriculum Learning.** Our framework incorporates a curriculum learning (CL) strategy that gradually increases training difficulty

| Model | Faithfulness | Novelty | Practicality | Coherence |
|---|---|---|---|---|
| Stable Diffusion | 3.71 | 3.78 | 3.37 | 3.37 |
| Kandinsky3 | 3.28 | 4.07 | 3.08 | 4.08 |
| ConceptLab | 3.33 | 4.57 | 3.03 | 4.04 |
| DALL-E | **4.08** | 4.32 | 3.32 | 4.56 |
| SYNTHIA (Ours) | 4.01 | **4.57** | **3.41** | **4.78** |

Table 2: **Evaluation Results with GPT-4o-mini model.**

during the fine-tuning process. To assess its effectiveness, we compare model performance with and without CL. Specifically, we train a base T2I model by randomly shuffling the training data and evaluate both settings using the absolute automatic evaluation protocol. As shown in the learning curve (Fig. 5), our CL approach leads to significantly higher scores even in the early stages of training compared to random training, highlighting its role in accelerating learning and guiding the model toward generating high-quality novel concepts. In contrast, training with CL results in a noticeable performance drop (Table 7), further demonstrating the importance of CL in our framework.

**Number of Positive Affordances.** To evaluate the impact of increasing the number of positive affordances in the input prompt, we conduct additional experiments using prompts with three and four affordances. As shown in Tables 4 and 5, all methods experience a slight performance drop as input complexity increases. However, SYNTHIA consistently maintains high scores in novelty and coherence, showing strong generalization to more complex affordance combinations—even though it is trained only on two affordance pairs. This result highlights the scalability of our method in generating high-quality, functionally coherent, and visually novel concepts without additional fine-tuning, showcasing its robustness and generalization capabilities. Moreover, SYNTHIA shows strong performance using only 600 affordance pairs with curriculum learning, accelerating training and enhancing generalization to novel affordance configurations.

**Evaluation Model Selection.** We further conducted experiments using different models, especially GPT-4o-mini, applying the same evaluation prompts and detailed criteria (Table 2). We observe consistent agreement in the results across different GPT-4o checkpoints, with our model consistently outperforming existing baselines across all metrics, even surpassing DALL-E model in novelty, practicality, and coherence, where DALL-E achieves 4.27, 3.21, and 4.52, respectively, when evaluated with GPT-4o-2024-08-06 (Faithfulness 4.02).

## 5 Conclusion

In this work, we take a step toward affordance-aware generative design by tackling a critical gap in existing text-to-image models—their inability to synthesize visually novel and functionally meaningful designs. With our model, SYNTHIA, we move beyond stylistic variation to purposeful creation, incorporating a hierarchical concept ontology with curriculum-guided fine-tuning to teach models how to compose affordances into unified, novel concepts. Our proposed four-dimensional evaluation metrics capture how well generated concepts align with real-world design principles. Through both human and automatic evaluations, SYNTHIA demonstrates clear superiority in designing imaginative but functionally grounded concepts. By bridging creativity with utility, this work advances AI-driven generative design.

## Limitations

Our work tackles an important yet underexplored problem of retaining functional coherence in AI for design using T2I models. While our model, in comparison to other state-of-the-art models, is able to generate more coherent and faithful images provided a set of affordances, e.g., `brew`, `cut` as in Fig. 1a, our work inherently relies on the human intuition to evaluate the novelty of the generated concepts. Although we try to alleviate the human bias and lack of coverage using LLM-as-a-judge for automatic evaluation, the question may persist. Moreover, although our concept ontology covers many different concept categories, it does not cover every plausible concept category in the real-world. It would be interesting to see follow-up works explore the direction of constructing a more diverse, richer concept ontology, which in turn would contribute to the generation of more novel concept designs.

## Acknowledgement

This research is based upon work supported by U.S. DARPA ECOLE Program No. #HR00112390060 and an Office of Naval Research (ONR) award #N00014-23-1-2780. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

## References

Vladimir Arkhipkin, Andrei Filatov, Viacheslav Vasilev, Anastasia Maltseva, Said Azizov, Igor Pavlov, Julia Agafonova, Andrey Kuznetsov, and Denis Dimitrov. 2023. Kandinsky 3.0 technical report. *Preprint*, arXiv:2312.03511.

Zhipeng Bao, Yijun Li, Krishna Kumar Singh, Yu-Xiong Wang, and Martial Hebert. 2024. Separate-and-enhance: Compositional finetuning for text2image diffusion models. *Preprint*, arXiv:2312.06712.

Alice Cai, Steven R Rick, Jennifer L Heyman, Yanxia Zhang, Alexandre Filipowicz, Matthew Hong, Matt Klenk, and Thomas Malone. 2023. Designaid: Using generative ai and semantic diversity for design inspiration. In *Proceedings of The ACM Collective Intelligence Conference*, pages 1–11.

Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. 2024. Pixart-$\sigma$: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *European Conference on Computer Vision*, pages 74–91. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. 2024. Scaling rectified flow transformers for high-resolution image synthesis. *Preprint*, arXiv:2403.03206.

G. Fauconnier and M. Turner. 2002. *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. Basic Books.

Ji Han, Feng Shi, Dongmyung Park, Liuqing Chen, and Peter R. N. Childs. 2018. The conceptual distances between ideas in combinational creativity. In *Proceedings of the DESIGN 2018 15th International Design Conference*.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. 2023. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*.

Hyung-Kwon Ko, Gwanmo Park, Hyeon Jeon, Jaemin Jo, Juho Kim, and Jinwook Seo. 2023. Large-scale text-to-image generation models for visual artists' creative works. In *Proceedings of the 28th international conference on intelligent user interfaces*, pages 919–933.

Gihyun Kwon, Simon Jenni, Dingzeyu Li, Joon-Young Lee, Jong Chul Ye, and Fabian Caba Heilbron. 2024. Concept weaver: Enabling multi-concept fusion in text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8880–8889.

David Chuan-En Lin, Hyeonsu B Kang, Nikolas Martelaro, Aniket Kittur, Yan-Ying Chen, and Matthew K Hong. 2025. Inkspire: Supporting design exploration with generative ai through analogical sketching. *arXiv preprint arXiv:2501.18588*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

Kevin Ma, Daniele Grandi, Christopher McComb, and Kosa Goucher-Lambert. 2023. Conceptual design generation using large language models. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, volume 87349, page V006T06A021. American Society of Mechanical Engineers.

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

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. *Preprint*, arXiv:2103.00020.

Tanzila Rahman, Shweta Mahajan, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, and Leonid Sigal. 2024. Visual concept-driven image generation with text-to-image diffusion model. *arXiv preprint arXiv:2402.11487*.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021a. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021b. Zero-shot text-to-image generation. *Preprint*, arXiv:2102.12092.

Harsh Rangwani, Aishwarya Agarwal, Kuldeep Kulkarni, R Venkatesh Babu, and Srikrishna Karanam. 2024. Crafting parts for expressive object composition. *arXiv preprint arXiv:2406.10197*.

Elad Richardson, Kfir Goldberg, Yuval Alaluf, and Daniel Cohen-Or. 2024. Conceptlab: Creative concept generation using vlm-guided diffusion prior constraints. *ACM Transactions on Graphics*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference* *on computer vision and pattern recognition*, pages 10684–10695.

B. Sankar and Dibakar Sen. 2024. A novel idea generation tool using a structured conversational ai (cai) system. *Preprint*, arXiv:2409.05747.

Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. 2024. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8543–8552.

Kihyuk Sohn, Nataniel Ruiz, Kimin Lee, Daniel Castro Chin, Irina Blok, Huiwen Chang, Jarred Barber, Lu Jiang, Glenn Entis, Yuanzhen Li, et al. 2023. Styledrop: Text-to-image generation in any style. *arXiv preprint arXiv:2306.00983*.

Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. 2024. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*, pages 1–18. Springer.

Yael Vinker, Andrey Voynov, Daniel Cohen-Or, and Ariel Shamir. 2023. Concept decomposition for visual exploration and inspiration. *ACM Transactions on Graphics (TOG)*, 42(6):1–13.

Arkhipkin Vladimir, Viacheslav Vasilev, Andrei Filatov, Igor Pavlov, Julia Agafonova, Nikolai Gerasimenko, Anna Averchenkova, Evelina Mironova, Bukashkin Anton, Konstantin Kulikov, Andrey Kuznetsov, and Denis Dimitrov. 2024. Kandinsky 3: Text-to-image synthesis for multifunctional generative framework. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 475–485, Miami, Florida, USA. Association for Computational Linguistics.

Ye Wang, Nicole B Damen, Thomas Gale, Voho Seo, and Hooman Shayani. 2024. Inspired by ai? a novel generative ai system to assist conceptual automotive design. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, volume 88407, page V006T06A030. American Society of Mechanical Engineers.

Zeyue Xue, Guanglu Song, Qiushan Guo, Boxiao Liu, Zhuofan Zong, Yu Liu, and Ping Luo. 2024. Raphael: Text-to-image generation via large mixture of diffusion paths. *Advances in Neural Information Processing Systems*, 36.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Yu Zhou, Bingxuan Li, Mohan Tang, Xiaomeng Jin, Te-Lin Wu, Kuan-Hao Huang, Heng Ji, Kai-Wei Chang, and Nanyun Peng. 2025. Contrastive visual data augmentation. *Preprint*, arXiv:2502.17709.

Qihao Zhu and Jianxi Luo. 2023. Generative transformers for design concept generation. *Journal of Computing and Information Science in Engineering*, 23(4):041003.

# Supplementary Materials

## A Hierarchical Concept Ontology Structure

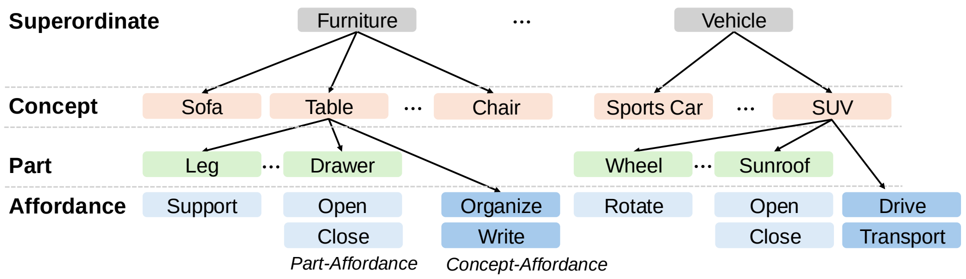


Figure 6: Hierarchical Concept Ontology.

## B Experiment Details

### B.1 Data Generation

To construct high-quality novel concept designs for each sampled affordance pair, we first generate candidate descriptions that satisfy the desired functionalities while ensuring visual novelty relative to existing concepts. Specifically, we prompt GPT-4o (OpenAI, 2024) to produce 10 diverse image captions for each affordance pair, describing novel concepts that fulfill the given affordances and differ in appearance from known objects. We then use these captions to generate 10 corresponding images using DALL-E (Ramesh et al., 2021b), resulting in a set of pseudo novel concepts.

Prompt for Image Caption Generation

You are a creative assistant who designs diverse novel concepts satisfying given conditions and generates a description of the concept. You should design three different novel concepts where each has all functions in the given positive constraints while the concept is different from the given negative constraints.
Generate three different descriptions of three novel concepts that contain visible unique characteristics to use generated descriptions as image captions to generate images. Each description should consist of at most three sentences and contain given positive constraints but should not contain non-visible characteristics such as sound, smell, and taste. You must not simply combine multiple existing concepts that have each function but creatively design a single concept that has multiple functions at once. Generate three descriptions of three novel concepts that are not similar to each other but distinct, and each description should be clear without unnecessary explanations for generating images. Please separate each description with \n\n. Simply follow the format given in the example below.

```
{
    Positive Constraints: [sit, store]
    Negative Constraints: [chair, car, sofa, bench, shelve, drawer]
    Image Captions: [“...”]
}
```

### B.2 Dataset

To evaluate the performance of our proposed method, we conduct our experiments by constructing a dataset from two types of resources:

- **Existing Concept Images**: For each existing concept in our ontology, we collect a dataset of 60 images from external platforms including Google Images and iStock. To ensure that the dataset is object-centric and minimizes noise, we filter out low-quality images using CLIP model (Radford et al., 2021). Specifically, we compute the similarity between the image embeddings and the text embeddings of the "`a photo of {concept name}`", selecting top-5 images with the highest similarity scores for each concept used as negative constraints.

- **Generated Novel Concept Images**: With our affordance sampling, we uniformly sample 600 affordance pairs among 82K possible pairs for fine-tuning. For the test dataset, we select 500 affordance pairs among the ones not used for fine-tuning. We use the generated images from the sampled affordances for fine-tuning and evaluation. The overall statistics can be found in Table 3.

| **Dataset** | Fine-tuning | | Inference |
|---|---|---|---|
| | Existing | Generated | Generated |
| **# Concepts** | 772 | 600 | 500 |
| **# Images** | 3860 | 1800 | 500 |

Table 3: Statistics of the datasets.

### B.3 Baseline Methods

- **Stable Diffusion** (Esser et al., 2024) is a strong baseline model for high-fidelity image synthesis, which is built on a diffusion-based framework. In this work, we leverage the pre-trained *stable-diffusion-3.5-large* model as the foundation model for the text-to-image task to generate a novel concept. Due to the limited context window length, we input only the positive constraints that contain the desired affordances, omitting the negative constraints associated with existing concepts.

- **Kandinsky** (Arkhipkin et al., 2023; Vladimir et al., 2024) model serves as another strong baseline model for comparison. We utilize the pre-trained *Kandinsky 3.0* model without any finetuning as the baseline, aligning it with the foundation model used in our proposed framework. This approach ensures a consistent starting point for evaluation and a fair comparison. This baseline allows us to effectively demonstrate the impact of our proposed training framework by comparing performance before and after the fine-tuning process.

- **ConceptLab** (Richardson et al., 2024) is a state-of-the-art framework designed for creative concept generation, which leverages an innovative approach that formulates the generation problem as an optimization process over the output space of the diffusion prior. It adopts a similar input format to the one used in our setting. We follow the finetuning process in the original framework, applying our training data to generate novel concepts. We then compare the generation quality during the inference time.

### B.4 Hyperparameter Settings

To fine-tune the diffusion model with the constructed dataset from DALL-E, we use the optimizer AdamW (Loshchilov and Hutter, 2019) with a learning rate of $\eta = \{5 * 10^{-6}, 10^{-6}, 5 * 10^{-7}\}$ to avoid the catastrophic forgetting problem. During the curriculum learning, we split the dataset into three groups based on their difficulty and trained 20 epochs for each group. For the weight factor in the triplet loss, we set $\gamma = \{0, 0.2, 0.5, 0.8, 1\}$. We finetune the UNet part in the pre-trained model and freeze the weights of other components in the diffusion model. The training takes about 1 GPU hour with NVIDIA A100 GPU.

## B.5 Automatic Evaluation

### B.5.1 Absolute Automatic Evaluation

Prompt for Absolute Automatic Evaluation

Please act as an impartial evaluator to assess the quality of a single concept image generated by an AI, based on the user's requirements. Your evaluation should use the following three criteria, each scored on a scale of 1 to 5:

Faithfulness: Evaluate how well the object aligns with the provided instructions, including its intended affordances and functionalities. Does the text and image together indicate that the object serves the purpose for which it was designed?
Scoring:
5: Flawlessly combines all specified functionalities as per the instructions. Text and image work in harmony to demonstrate a well-designed and fully functional object.
4: Fulfills most instructions and intended functionalities, with only minor inconsistencies or missing details. The text and image are mostly aligned.
3: Partially fulfills the instructions. Some functionalities are present but not well-integrated or consistent. There may be a minor mismatch between text and image.
2: Struggles to meet the provided instructions, missing key functionalities or combining them poorly. Text and image may conflict.
1: Does not follow the instructions at all. Functionalities are completely missing, irrelevant, or nonsensical.

Novelty: Assess the originality and innovation of the design. Does the object show an exciting, novel design that would surprise or intrigue users?
Scoring:
5: Highly innovative, unique, and impressive. Inspires curiosity or excitement, making it highly desirable to explore.
4: Contains interesting and novel elements, showing clear creative thought and appeal.
3: Displays moderate novelty, with some unique features but remaining relatively conventional or uninspiring.
2: Shows limited novelty, with minimal creativity and overly simplistic or derivative design.
1: Entirely unoriginal and mundane, lacking any creativity and appearing common or uninspiring.

Practicality: Evaluate the real-world applicability of the object. Does the design make sense for human use? Would it align with human preferences and be feasible for production?
Scoring:
5: Extremely practical and human-centric. Highly functional, aligns perfectly with human preferences, and seamlessly fits into real-world scenarios.
4: Mostly practical and applicable, with minor limitations that could be addressed to improve usability.
3: Somewhat practical but with notable flaws or unrealistic elements that may limit usability in real-world scenarios.
2: Largely impractical, with significant flaws or inconsistencies that make it unlikely to be useful.
1: Entirely impractical and unusable, failing to align with human preferences or real-world feasibility.

Coherence: This metric evaluates whether the image generated by the AI model contains only one primary object as instructed, focusing on the object's clarity and functionality without the presence of additional, unintended objects.

Scoring:
5: The image perfectly showcases one distinct object that aligns with the described attributes. There are no extraneous objects or elements that distract from the main object.
4: The primary object is clear and mostly isolated, but there may be minor elements in the background or periphery that do not significantly detract from the main object.
3: The main object is present and identifiable, but there are other elements in the image that somewhat distract from its clarity and functionality.
2: The image contains multiple objects where the main object is not clearly dominant or distinguishable from other unnecessary elements.
1: The image primarily features multiple objects, making it difficult to identify the intended single object; the composition is cluttered or entirely irrelevant to the instruction.
Provide a score for each criterion, followed by a concise explanation justifying your ratings. Your final response should strictly follow this format:

```
{
    "Faithfulness": [Your Faithfulness Score],
    "Novelty": [Your Novelty Score],
    "Practicality": [Your Practicality Score],
    "Coherence": [Your Coherence Score]
}
```

### B.5.2 Relative Automatic Evaluation

Prompt for Relative Automatic Evaluation

Please act as an impartial evaluator to assess the quality of concept images generated by two AI concept generators based on the user's requirements. The evaluation criteria are as follows: Faithfulness: Evaluate how well the object aligns with the provided instructions, including its intended affordances and functionalities. Does the text and image together indicate that the object serves the purpose for which it was designed? Novelty: Assess the originality and innovation of the design. Does the concept demonstrate a surprising or intriguing approach that stands out as fresh and exciting? Practicality: Evaluate the real-world applicability of the concept. Does the design make sense for human use, align with user preferences, and appear feasible for production? Coherence: This metric evaluates whether the image generated by the AI model contains only one primary object as instructed, focusing on the object's clarity and functionality without the presence of additional, unintended objects. Provide your answer based on the following available choices: "A" if the first image is better, "B" if the second image is better, "C" if both are equally strong. Your final response should strictly follow this format:

```
{
    "Faithfulness": [Your Faithfulness Choice],
    "Novelty": [Your Novelty Choice],
    "Practicality": [Your Practicality Choice],
    "Coherence": [Your Coherence Choice]
}
```

### B.6 Human Evaluation

#### Human Evaluation Instruction

**Objective**
The goal of this evaluation is to assess the quality of novel concepts that integrate multiple affordances into a single, coherent design. Affordance refers to the functional properties of an object or its components. For example, a sofa affords the function of sitting, while its legs provide the function of support.
As an annotator, you will evaluate the given concepts based on four key metrics: **faithfulness**, **novelty**, **practicality**, and **coherence**. Each metric is defined below, along with its respective scoring criteria.

**Evaluation Criteria**
Faithfulness (Does the concept effectively integrate the specified affordances?)
This metric assesses whether the generated concept successfully incorporates all provided affordances in a meaningful and functional manner.

Scoring Scale:
5 – Fully integrates all specified affordances, demonstrating a well-designed and fully functional object.
4 – Incorporates all affordances with minor inconsistencies or slight missing details.
3 – Partially fulfills the affordances; some functionalities are present but not well-integrated or consistent.
2 – Struggles to meet the provided affordances; key functionalities are missing or poorly combined.
1 – Does not incorporate the specified affordances; functionalities are entirely missing, irrelevant, or nonsensical.

Novelty (To what extent does the concept demonstrate originality and innovation?)
This metric evaluates the uniqueness and creative appeal of the design, considering whether it introduces novel elements that would intrigue or surprise users.
Scoring Scale:
5 – Highly innovative and unique; presents a strikingly original concept that is engaging and thought-provoking.
4 – Contains clear novel elements, demonstrating creative thought and originality.
3 – Moderately novel; some unique aspects are present, but the overall concept remains relatively conventional.
2 – Limited novelty; the design appears simplistic, derivative, or lacking in creativity.
1 – Entirely unoriginal and uninspiring, closely resembling existing designs with no innovative aspects.

Practicality (Is the design feasible and suitable for real-world use?)
This metric assesses whether the concept is functionally viable and aligned with human preferences and usability considerations.
Scoring Scale:
5 – Highly practical and user-centered; seamlessly functional and feasible for real-world applications.
4 – Mostly practical; minor limitations exist but do not significantly hinder usability.
3 – Somewhat practical; contains notable flaws or unrealistic elements that may limit real-world applicability.
2 – Largely impractical; significant design flaws make real-world usability unlikely.

1 – Entirely impractical and non-functional; does not align with human preferences or feasibility constraints.
Coherence (Does the image clearly depict a single, distinct object?)
This metric evaluates whether the design presents a singular, well-defined object, free from extraneous elements that may obscure its intended functionality.
Scoring Scale:
5 – The image clearly and exclusively depicts a single object that integrates all specified affordances without any distractions.
4 – The primary object is distinct and well-defined, though minor background elements may be present without significantly detracting from clarity.
3 – The main object is identifiable, but additional elements in the image introduce some visual or conceptual distractions.
2 – The image contains multiple objects, making it difficult to distinguish the intended primary object or missing at least one affordance.
1 – The image primarily features multiple objects, with affordances spread across different elements rather than a unified concept, making it unclear what the primary object is.

**Final Instructions**
Please evaluate each concept independently based on the above criteria.
Assign a score for each metric according to the provided descriptions.
If a concept does not fit neatly into the scoring categories, use your best judgment to determine the most appropriate score.
Your evaluations will contribute to assessing the effectiveness of novel concept generation and help improve future designs. Thank you for your participation.

# C Additional Experimental Results

## C.1 Results of Relative Automatic Evaluation

As described in Sec 4.3, we perform relative evaluation following the protocol B.5.2, comparing each method s generated concepts against those from DALL-E. We show that SYNTHIA achieves higher win rates than the baselines in all metrics when evaluated against DALL-E. Moreover, we conduct direct pairwise comparisons between SYNTHIA and the baselines. As shown in Fig. 7, SYNTHIA outperforms all baselines across every metric, demonstrating its superior ability to design novel concepts that are both visually novel and functionally coherent, given a set of desired affordances.

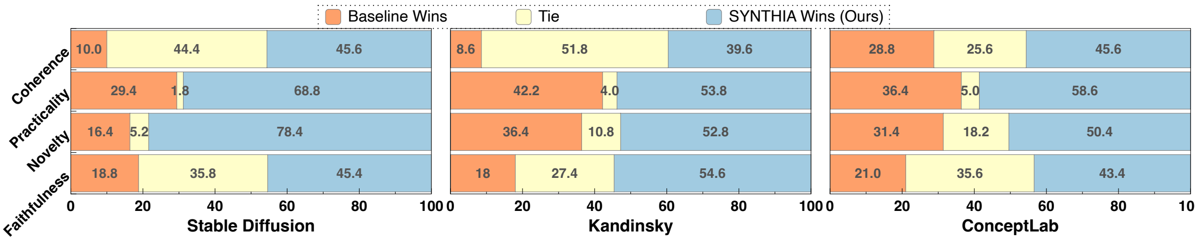


Figure 7: **Results of the relative automatic evaluation.** We compare the quality of concepts generated from our models with ones from existing T2I models. Numbers indicate the percentage (%) of baseline model wins, ties, and SYNTHIA model wins.

### C.2 Results of Human Evaluation

In addition to the absolute inter-annotator agreement (IAA) reported in Sec 4.3, we further visualize the agreement between individual human annotators. As shown in Fig. 8, the overall agreement is high, supporting the reliability and validity of our human evaluation results.

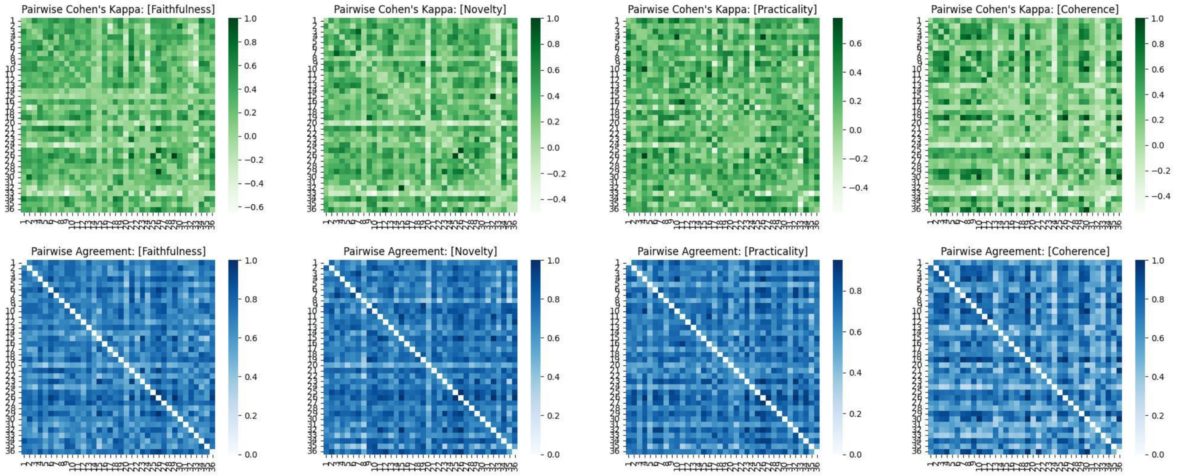


Figure 8: Results of the human evaluation.

### C.3 Effectiveness of Affordance-based Textual Prompt

During the inference for novel concept generation, we align textual prompts with affordances as we described in Sec 3.3. To further demonstrate the effectiveness of incorporating affordance to guide the novel concept synthesis, we additionally conduct experiments with textual input based on concept-level prompts without adding additional signals of affordances. We especially use prompts of "*a new design that includes functions from {concepts that possess desired affordances – e.g., sofa, chair}.*". For a fair comparison, we use the same fine-tuned models and baseline models but only utilize different prompts (concept-level prompt vs. affordance-level prompt) to evaluate the quality of the generated concepts. As shown in Table 4, affordance-level prompts facilitate a better design across different models, achieving an improvement in faithfulness and coherence. This demonstrates that aligning textual prompts with affordances can benefit even off-the-shelf T2I models to achieve better functionally coherent concepts. However, they are still far behind our models, showing the fundamental challenges in affordance-aware novel concept synthesis.

| Model | Type | Faithfulness | Novelty | Practicality | Coherence |
|---|---|---|---|---|---|
| Stable Diffusion | Concept | 2.99 | 3.83 | 2.87 | 2.65 |
| | Affordance | 3.73 | 3.80 | 3.33 | 3.97 |
| Kandinsky3 | Concept | 3.10 | 3.92 | 2.79 | 3.24 |
| | Affordance | 3.52 | 3.91 | 3.11 | 4.20 |
| ConceptLab | Concept | 3.61 | 4.13 | 3.14 | 4.31 |
| | Affordance | 3.78 | 4.03 | 3.08 | 4.27 |
| SYNTHIA (Ours) | Concept | 3.90 | 4.34 | 3.28 | 4.60 |
| | Affordance | 3.99 | 4.45 | 3.36 | 4.76 |

Table 4: **Results of the automatic evaluation with three positive affordances.** We compare our method with baselines. For each metric, a higher number indicates a better performance, where the score ranges between 0 and 5.

| Model | Type | Faithfulness | Novelty | Practicality | Coherence |
|---|---|---|---|---|---|
| Stable Diffusion | Concept | 2.74 | 3.73 | 2.73 | 2.38 |
| | Affordance | 3.41 | 3.82 | 3.08 | 3.71 |
| Kandinsky3 | Concept | 2.92 | 3.88 | 2.62 | 2.82 |
| | Affordance | 3.33 | 3.87 | 3.13 | 4.07 |
| ConceptLab | Concept | 3.67 | 4.03 | 3.03 | 4.42 |
| | Affordance | 3.61 | 3.98 | 2.98 | 4.17 |
| SYNTHIA (Ours) | Concept | 3.85 | 4.36 | 3.14 | 4.59 |
| | Affordance | 3.86 | 4.52 | 3.25 | 4.80 |

Table 5: **Results of the automatic evaluation with four positive affordances**. We compare our method with baselines. For each metric, a higher number indicates a better performance, where the score ranges between 0 and 5.

### C.4 Effectiveness of Affordance Sampling

| Model | Type | Close | Distant |
|---|---|---|---|
| Stable Diffusion | Concept | 3.92 | 4.13 |
| | Affordance | 3.71 | 3.88 |
| Kandinsky3 | Concept | 4.04 | 4.13 |
| | Affordance | 4.18 | 4.14 |
| ConceptLab | Concept | 4.18 | 4.17 |
| | Affordance | 3.96 | 4.20 |
| SYNTHIA (Ours) | Concept | 4.26 | 4.33 |
| | Affordance | 4.46 | 4.59 |

Table 6: **Results of the automatic novelty evaluation with different distances.** We compare our method with baselines. For each metric, a higher number indicates a better performance, where the score ranges between 0 and 5.

### C.5 Effectiveness of Curriculum Learning

| Model | Type | Faithfulness | Novelty | Practicality | Coherence |
|---|---|---|---|---|---|
| SYNTHIA without CL | Concept | 3.61 | 4.03 | 3.07 | 4.30 |
| | Affordance | 3.56 | 4.13 | 2.95 | 4.59 |
| SYNTHIA (Ours) | Concept | 3.97 | 4.33 | 3.30 | 4.51 |
| | Affordance | 3.99 | 4.55 | 3.35 | 4.81 |

Table 7: **Results of the automatic evaluation with and without curriculum learning.** For each metric, a higher number indicates a better performance, where the score ranges between 0 and 5.

### C.6 More Visual Examples

We include additional visual examples comparing the baseline methods with SYNTHIA (Ours). As illustrated in Figure 9, our model outperforms in generating concepts that align closely with the input affordance pairs. Moreover, SYNTHIA consistently produces outputs with higher quality across 4 key dimensions, including faithfulness, novelty, practicality, and coherence.

## D Relevance to NLP domain

While our work focuses on Text-to-Image (T2I) generation, it is fundamentally grounded in Natural Language Processing (NLP) domain. Specifically, we leverage NLP methods to extract concept-affordance

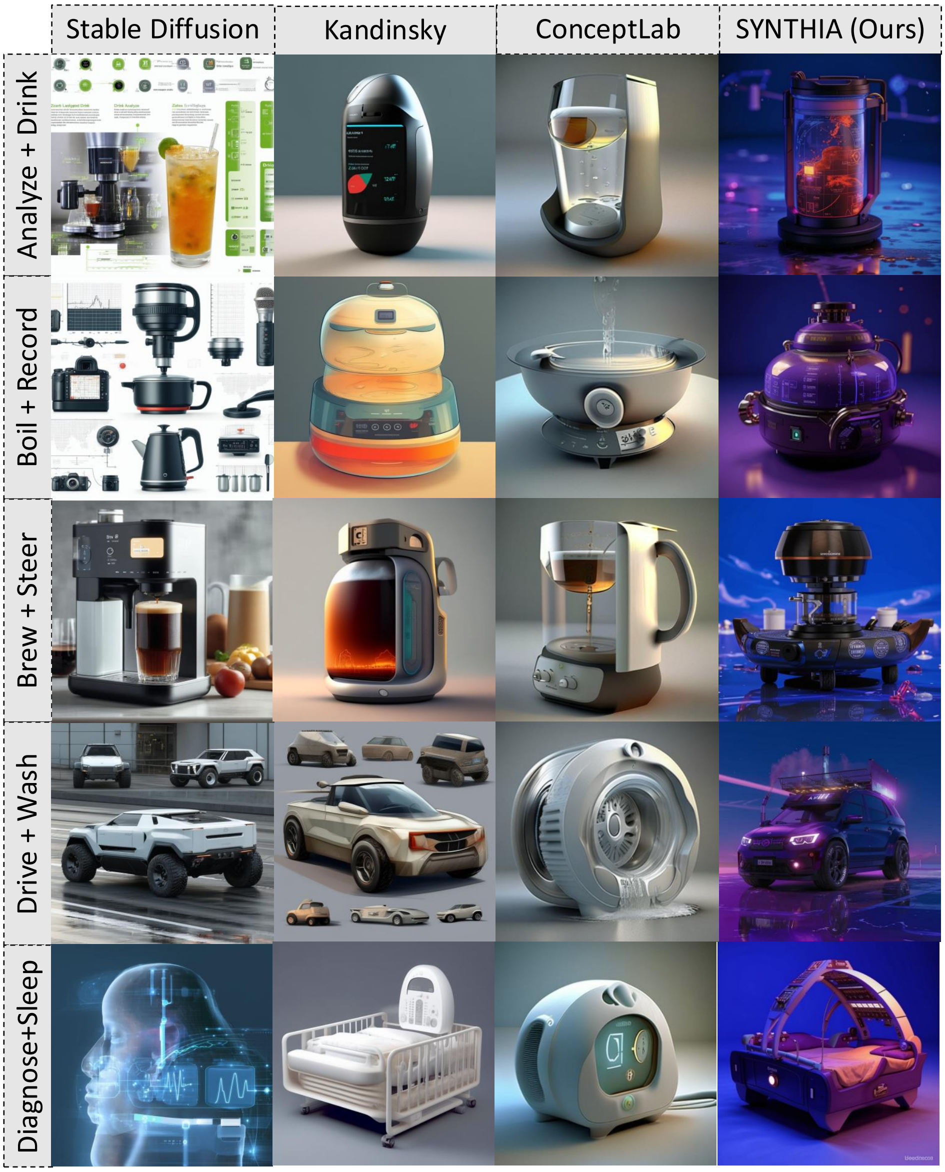


Figure 9: More generated examples from the baseline methods and SYNTHIA (ours).

associations from commonsense knowledge bases and incorporate the structured knowledge into the image generation process. Our interdisciplinary approach not only expands the horizons of text-to-image generation, but also advances core NLP research by demonstrating how linguistic and semantic understanding can inform and enrich multimodal systems. This contribution aligns closely with the NLP community growing interest in multimodal AI and the integration of language with other modalities.

## Ethical Consideration

We acknowledge that our work is aligned with the *ACL Code of the Ethics* [1] and will not raise ethical concerns. We do not use sensitive datasets/models that may cause any potential issues/risks.

[1] https://www.aclweb.org/portal/content/acl-code-ethics