Bridging the Sensory Gap: Visual Injection for Taxonomy Completion

Yuhang Niu[♡], **Hongyuan Xu**[♡],
Ciyi Liu, **Bofan Wei**, **Jiaqi Ye**, **Yanlong Wen**[♢], **Xiaojie Yuan**
College of Computer Science, Nankai University, Tianjin, China
{niuyuhang,xuhongyuan,liuciyi,weibofan}@dbis.nankai.edu.cn
{wenyl,yuanxj}@nankai.edu.cn

Abstract

Taxonomy Completion aims to automatically integrate new concepts into existing hierarchies. However, existing text-only methods suffer from a "Sensory Gap": they struggle to differentiate ambiguous definitions (e.g., Latte vs. Cappuccino) and miss visual grouping signals. Consequently, they often misinterpret lexical overlaps as hierarchical dependencies, leading to erroneous structural predictions. To bridge this, we propose VITC, a framework leveraging Visual Injection for Taxonomy Completion. By mapping synthesized images into intrinsic *pseudo-tokens*, we enable the text encoder to perform holistic structural reasoning. To address injection challenges, we introduce Adaptive Residual Fusion, which decouples magnitude from selection to prevent visual signals from being drowned out, and the Multimodal Guided Adaptive Reweighting strategy, which leverages cross-modal consensus (Mutual Rescue and Complementary Mining) to filter noise and identify hard negatives. Experiments on three datasets demonstrate that VITC achieves state-of-the-art performance, delivering an average absolute gain of over 19% in Hit@1. Code is available at https://github.com/nyh-a/VITC.

1 Introduction

Taxonomies organize concepts into hierarchical hypernym-hyponym ("*is-a*") structures, serving as the structural backbone for knowledge-driven applications, ranging from recommendation systems (Zhang et al., 2014) to enhancing Large Language Models' reasoning (Pan et al., 2024). However, static taxonomies lag behind emerging terms, necessitating automatic **Taxonomy Completion (TC)** to integrate new query concepts into existing hierarchies. As illustrated in Figure 1, the goal is to insert the query "*Cappuccino*" into its optimal position: *Espresso Drink (Parent)* → *Cappuccino (Query)* → *Dry Cappuccino (Child)*.

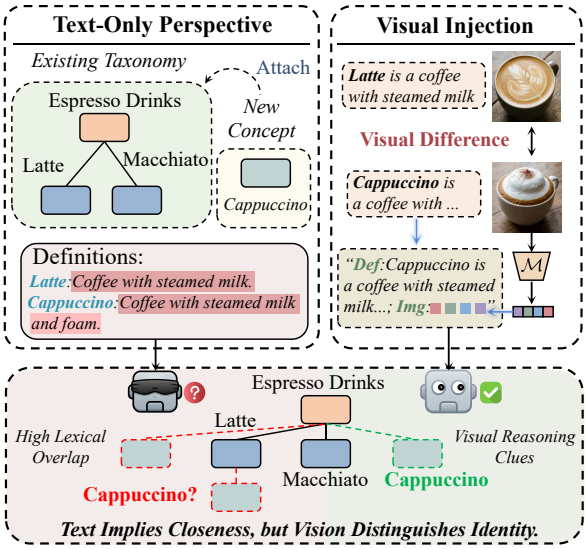

Figure 1: Illustration of the **"Sensory Gap"** in Taxonomy Completion using the "Cappuccino" example. **(Left)** High lexical overlap between "Latte" and "Cappuccino" misleads text-only models into inferring a false Parent-Child hierarchy. **(Right)** Visual Injection introduces distinctive cues (e.g., foam) via pseudo-tokens. **(Bottom)** Visual grounding resolves the ambiguity, correctly identifying them as mutually exclusive Siblings.

Existing literature typically approaches the task from two perspectives: *structural* and *semantic* (Xu et al., 2025b). While structural methods leverage topology (Liu et al., 2021; Shen et al., 2020), state-of-the-art methods mainly rely on Language Models (LMs) to capture fine-grained semantic nuances (Niu et al., 2024; Xu et al., 2023). Since standard taxonomic benchmarks are purely textual, these methods are inherently text-only. This unimodal reliance creates a fundamental **"Sensory Gap"**. First, they struggle with *Descriptive Ambiguity*. As illustrated in Figure 1, the lexical overlap between "*Latte*" and "*Cappuccino*" often misleads text-only models into inferring a false Parent-Child relationship. Visual cues (e.g., texture) are essential to correct this, distinguishing them as mutually

[♡] Equal contribution. [♢] Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6092–6107

exclusive Siblings. Second, they miss *Visual Reasoning Clues*. Concepts like "*Bagel*" and "*Donut*" may differ textually but share strong visual resemblances (e.g., shape). This serves as an explicit signal to cluster them under the same parent, a connection that structure-free text often overlooks.

To bridge this "Sensory Gap", we can leverage the rich world knowledge of modern generative models (e.g., DALL-E 3, Betker et al., 2023) to synthesize images for concepts. However, effectively integrating these signals goes beyond simple feature concatenation. TC is fundamentally a structural matching task that evaluates the compatibility between a *Query* and a topological *Position*. Traditional late-fusion architectures are ill-suited here because they encode images and text in isolation (Baldrati et al., 2022; Wen et al., 2024). This fractures the reasoning process: the text encoder cannot "consult" the visual appearance to verify its semantic interpretation. To address this, we propose **VITC**, which employs a **Generative Visual Injection** strategy. By synthesizing representative images for concepts and mapping them into *pseudo-tokens*, we integrate visual features directly into the input text sequence. This treats visual features as intrinsic textual units, allowing the LM to "see" while it "reads". This enables deep *Mutual Grounding*: the textual context dynamically attends to relevant visual cues to resolve ambiguity (e.g., checking for foam), while visual tokens anchor abstract definitions to concrete sensory evidence.

Despite its benefits, the injection introduces a **Density-Selectivity Dilemma**. Since we map images into a compact sequence ($k \approx 4$) to fit the LM's context, visual signals are numerically "drowned out" by lengthy textual definitions ($N \geq 50$) during aggregation. Standard gating (Srivastava et al., 2015) fails here: a high gate value is needed to boost visibility (prevent drowning), but a low value is required to filter out irrelevant visual cues. To resolve this, we propose **Adaptive Residual Fusion**, which *decouples magnitude from selection*. A learnable scalar acts as a "volume knob" to *boost* the visual signal's magnitude, ensuring it survives the textual ocean, while a separate semantic gate acts as a "switch" to *verify* its relevance.

Finally, to optimize VITC for discriminative structural matching, we employ a contrastive learning framework (Gao et al., 2021; Niu et al., 2024). However, the integration of generated images introduces a **Data Quality Dilemma**, acting as a *double-edged sword* for contrastive training. On one hand, hallucinations introduce *noise* (risking data validity); on the other, visual resemblance between siblings offers valuable *hard negatives* (Robinson et al., 2021) (improving discriminability). Standard strategies fail to distinguish these cases, treating both as simple negatives or outliers. To govern this, we propose **Multimodal Guided Adaptive Reweighting (MGAR)**, which leverages cross-modal consensus: (1) *Mutual Rescue (Handling Noise):* Conventional filters (Solatorio, 2024) often discard samples based on low textual similarity alone. MGAR acts as a "safety net", flagging a sample as noise only if *both* text and vision reject it, thus preserving valid long-tail concepts that are obscure in one modality. (2) *Complementary Mining (Handling Hard Negatives):* Visual similarity is symmetric and confusing for siblings. We turn this confusion into a training signal: if *either* modality flags a negative candidate as highly similar, we treat it as a hard negative and boost its penalty. This forces the model to learn a rigorous decision boundary, distinguishing symmetric visual likeness from asymmetric hierarchical entailment. Our main contributions are summarized as follows:

- We propose **VITC**, a framework utilizing **Visual Injection** for TC. By mapping images into *pseudo-tokens*, we bridge the "Sensory Gap" and enable the text encoder to verify semantic relations via deep visual grounding.
- We introduce **Adaptive Residual Fusion**, resolving token imbalance by *decoupling magnitude from selection*, and **MGAR**, optimizing training via cross-modal consensus (Mutual Rescue & Complementary Mining).
- Experiments on three datasets demonstrate state-of-the-art performance, achieving an average absolute gain of over **19%** in Hit@1 compared to purely textual baselines.

2 Related Work

Taxonomy Expansion and Completion. Automatic taxonomy enrichment is generally categorized into expansion and completion. Taxonomy expansion (TE) (Shen et al., 2020) anchors new concepts into an existing hierarchy but is typically constrained to leaf positions. To overcome this limitation, taxonomy completion (TC) (Zhang et al., 2021) allows insertions at any internal node. Subsequent work has advanced TC along several directions: to capture local topological context, Tax-

oEnrich (Jiang et al., 2022) and QEN (Wang et al., 2022) explicitly aggregate sibling information to refine position representations; GenTaxo (Zeng et al., 2021) and ICON (Shi et al., 2024) generate new concepts based on existing taxonomies; shifting focus to representation learning, TaxBox (Xue et al., 2024) leverages geometric box embeddings to model hypernymy as spatial containment, while CoSTC (Niu et al., 2024) employs contrastive learning with hard negative mining to distinguish subtle semantic boundaries. Building upon the exploitation of pre-trained language models (PLMs) (Liu et al., 2021; Xu et al., 2023, 2025a) to extract hierarchical knowledge, recent approaches like COMI (Xu et al., 2025b) scale up to leverage LLMs to jointly model fine-grained concept semantics and hierarchical structural dependencies. Despite these advances, existing methods remain inherently *text-dependent*, suffering from the "Sensory Gap": they struggle to differentiate concepts with high lexical overlap that possess distinct visual features. In contrast, VITC introduces Generative Visual Injection via pseudo-tokens, providing the encoder with explicit visual cues to resolve textual ambiguities that prior text-only baselines miss.

Textual Inversion. Textual Inversion (TI) (Gal et al., 2023) projects visual concepts into the embedding space of Language Models (LMs) as "pseudo-words". This "Image-as-a-Word" approach preserves the LM's inherent compositional reasoning without requiring complex cross-modal fusion architectures (Zhou et al., 2022; Voynov et al., 2023). Existing methods generally fall into two categories: *optimization-based inversion* (Gal et al., 2023; Cohen et al., 2022), which requires costly iterative updates, and the more efficient *prediction-based inversion*. Recent works in the latter category, such as Pic2Word (Saito et al., 2023) and iSEARLE (Agnolucci et al., 2025), employ dedicated mapping networks for direct feature projection. In this paper, we leverage the *prediction-based* strategy for efficient visual injection, repurposing visual tokens to aid discriminative structural reasoning rather than image generation.

Generative Synthesis vs. Retrieval. An alternative is to retrieve web images per concept (Zhu et al., 2023). We choose generative synthesis for three reasons. (1) *Polysemy resolution*: retrieval engines key on surface names (e.g., "Bank") and return mixed senses, whereas generators conditioned on the full definition produce images aligned with the specific taxonomic sense. (2) *Abstract coverage*: retrieval fails for abstract nodes (e.g., verbs like "Think"), while generative models yield consistent visual metaphors that serve as valid discriminative signals. (3) *Domain consistency*: retrieved images exhibit large stylistic variance (photos, cliparts, diagrams), whereas synthesized images follow a uniform style controlled by the prompt, facilitating pattern recognition for the vision encoder.

3 Methodology

3.1 Problem Formulation

A **taxonomy** is a directed acyclic graph $\mathcal{T} = (\mathcal{N}, \mathcal{E})$, where node $n \in \mathcal{N}$ represents a concept and edge $\langle n_p, n_c \rangle \in \mathcal{E}$ denotes a hypernym-hyponym relation. Given an existing taxonomy $\mathcal{T}^0$ and a set of new query concepts $\mathcal{C}$, the **Taxonomy Completion** task is to insert each query $q \in \mathcal{C}$ into $\mathcal{T}^0$ by identifying its optimal insertion position pos, defined by a candidate parent-child pair $\langle p, c \rangle$ where $p, c \in \mathcal{N}^0$. We formulate this as a metric learning problem: learning a scoring function $f(q, pos; \Theta)$ to measure the semantic compatibility between q and candidate position pos.

3.2 Framework Overview

To bridge the "Sensory Gap", VITC employs a **Generative Visual Injection** strategy. We first address the inherent lack of visual data in standard taxonomic benchmarks via **Visual Imputation**, utilizing DALL-E 3 (Betker et al., 2023) to synthesize canonical images for all nodes based on their definitions (executed as an offline pre-processing step, see Appendix A.4 for details). As illustrated in Figure 2, VITC operates in two stages: (1) **Structure-Aware Visual Mapping** (Sec. 3.3), which pretrains a map network to project synthesized images into intrinsic *pseudo-tokens*; and (2) **Deep Injection & Adaptive Fusion** (Sec. 3.4), which embeds these tokens into structural templates for **discriminative structural matching**, optimized via the consensus-driven **MGAR** strategy (Sec. 3.5).

3.3 Structure-Aware Visual Mapping

The goal of this stage is to construct a mapping interface that translates generated visual representations into the text encoder's input token space.

Visual Projection. Given a concept q with image I_q, we first extract the visual embedding $\mathbf{v}_q \in \mathbb{R}^{d_v}$ using a frozen vision encoder E_{vis} from a Vision-Language Model (e.g., BLIP (Li et al., 2022)). To

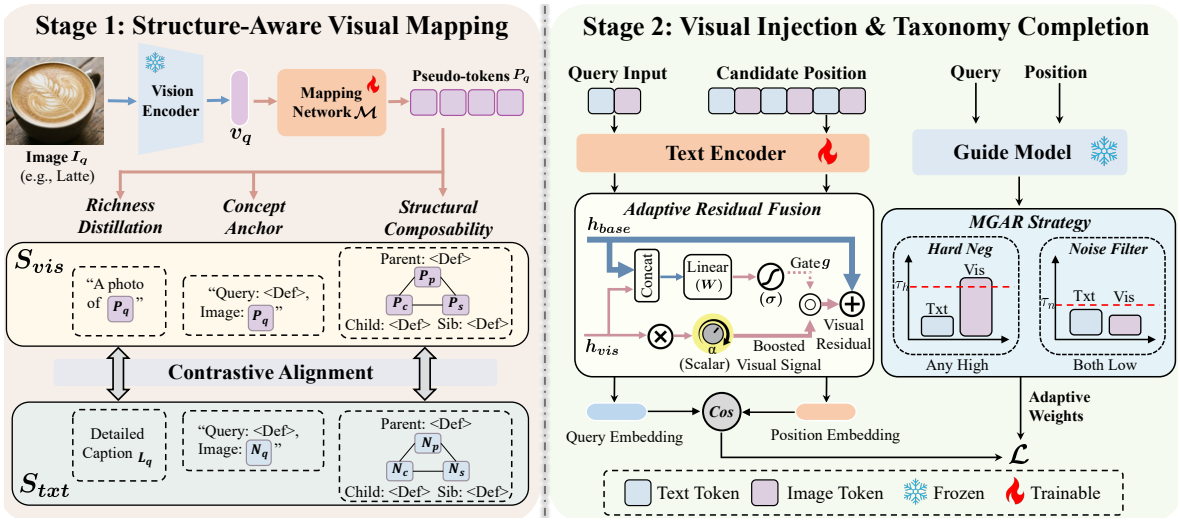

Figure 2: Overview of VITC. The framework includes Stage 1 for structure-aware visual mapping and Stage 2 for deep injection. Key components include Adaptive Residual Fusion for multimodal integration and the MGAR Strategy for robust optimization.

bridge the modality gap, we employ a lightweight mapping network $\mathcal{M}$ (a multi-layer perceptron) to project $\mathbf{v}_q$ into a sequence of k *pseudo-tokens* in the text embedding space:

$$\mathbf{P}_q = \mathcal{M}(\mathbf{v}_q) = \{p_1, \ldots, p_k\} \in \mathbb{R}^{k \times d_t} \quad (1)$$

Unlike optimization-based inversion that updates embeddings per image, $\mathcal{M}$ is a prediction network trained to directly generate sequence-compatible tokens that can interact with textual prompts.

Structure-Aware Pre-training. We optimize $\mathcal{M}$ by aligning the embedding of a visually-injected input sequence ($\mathcal{S}_{vis}$) with a text-only target sequence ($\mathcal{S}_{txt}$) via contrastive loss. To ensure the tokens capture both visual richness and structural role (enabling the "Mutual Grounding" described in Sec. 1), we design three proxy tasks:

Task A: Richness Distillation. To compel tokens to retain fine-grained visual details (e.g., texture, shape), we force them to reconstruct the semantics of the detailed image caption L_q (generated by DALL-E 3). The target is the raw caption $\mathcal{S}_{txt} = L_q$, while the input is constructed with pseudo-tokens: $\mathcal{S}_{vis} =$ "*A photo of* <$\mathbf{P_q}$>".

Task B: Concept Alignment. To prevent semantic drift (Baldrati et al., 2023), we anchor the tokens to the concept's definition and name using a template $\mathcal{T}_{\text{query}} =$ "*Query Node: Def: <D>, Image: <X>*", where D is the concept definition. The target $\mathcal{S}_{txt}$ fills <X> with the concept name N_q, while the input $\mathcal{S}_{vis}$ replaces it with pseudo-tokens.

Task C: Structural Composability. To adapt visual tokens to the hierarchical topology and enable "Mutual Grounding", we define a triplet template $\mathcal{T}_{pos}$ encompassing Parent, Child, and Sibling roles. We construct $\mathcal{S}_{vis}$ by replacing the names of *all three nodes* with their respective visual tokens. This compels the pseudo-tokens to retain semantic robustness even when embedded amidst multiple long definitions, simulating the *holistic structural reasoning* scenario:

$$\begin{aligned} \mathcal{S}_{txt} &= \mathcal{T}_{pos}(N_p, N_c, N_s) \\ \mathcal{S}_{vis} &= \mathcal{T}_{pos}([\mathbf{P}_p], [\mathbf{P}_c], [\mathbf{P}_s]) \end{aligned} \quad (2)$$

where $\mathcal{T}_{pos}$ concatenates the contexts: "*Parent Node: Def: <D_p>, Image: <X_p>; Child Node: Def: <D_c>, Image: <X_c>; Sibling Node: Def: <D_s>, Image: <X_s>*".

Anti-Shortcut Strategy. To prevent the mapping network from ignoring visual signals when definitions D are explicit, we apply *Dual-Level Dropout*: (1) discarding the entire definition field with probability ρ_{def}; and (2) randomly masking discrete word tokens within D. This forces $\mathbf{P}_q$ to encode essential semantics to compensate for missing text, ensuring the model learns to "consult" the visual tokens rather than treating them as redundant noise.

3.4 Deep Injection and Adaptive Fusion

Stage 2 executes the taxonomy completion task by incorporating the trained mapping network into a unified metric learning pipeline for discriminative structural matching. We encode the query q and

candidate position $pos = \langle p, c, s \rangle$ into multimodal embeddings $\mathbf{h}^{(q)}$ and $\mathbf{h}^{(pos)}$, followed by a specialized fusion strategy to resolve density imbalance.

Deep Injection Mechanism. Unlike disjoint late-fusion architectures (Baldrati et al., 2022), we construct a unified input to enable deep cross-modal interaction. Reusing Stage 1 templates, we encode q via the *identity template* ($\mathcal{T}_{query}$) and pos via the *Triplet Template* ($\mathcal{T}_{pos}$). Since the pseudo-tokens $\mathbf{P}$ are integrated into the sequence, they participate in the LM's global self-attention. This allows the textual context to dynamically "consult" relevant visual cues (e.g., checking for foam) to verify semantic alignment before the final aggregation.

Adaptive Residual Fusion. Standard Mean Pooling faces a density-selectivity dilemma. The highly condensed visual tokens ($k \approx 4$) are numerically "drowned out" by lengthy textual definitions ($N \geq 50$). A standard gate faces a conflict here: it must be *high* to boost the visual signal's visibility against the text, but *low* to filter out noise from low-quality generated images. To resolve this, we *decouple feature magnitude from semantic relevance*. Let h_{base} be the standard pooling of the full sequence, and h_{vis} be that of only the visual tokens. We formulate the final representation $\boldsymbol{h}_{final}$ as a learnable residual increment:

$$\boldsymbol{h}_{final} = \underbrace{\boldsymbol{h}_{base}}_{\text{Standard View}} + \underbrace{\boldsymbol{g} \odot (\boldsymbol{\alpha} \cdot \boldsymbol{h}_{vis})}_{\text{Visual Residual}} \quad (3)$$

This mechanism decouples the roles via two components: (1) *Visual Boosting ($\boldsymbol{\alpha}$)*: A learnable scalar acting as a "volume knob" that unconditionally *boosts* the visual magnitude to a level comparable with textual features, ensuring the signal is visible to the model; (2) *Relevance Verification ($\boldsymbol{g}$)*: A context-aware gate acting as a "switch" ($\mathbf{g} = \sigma(W_g[\boldsymbol{h}_{base}; \boldsymbol{h}_{vis}] + b_g)$). This gate dynamically verifies semantic relevance, filtering out visual noise when the generated image is unreliable. Finally, the matching score is computed via cosine similarity: $f(q, pos) = \cos(\mathbf{h}^{(q)}_{final}, \mathbf{h}^{(pos)}_{final})$.

3.5 Multimodal Guided Adaptive Reweighting

To optimize discriminative structural matching under **Data Quality Uncertainty** of generated images, we employ a robust contrastive learning strategy. Instead of treating all samples equally, we introduce **MGAR** to govern the trade-off between *Data Validity* (filtering noise) and *Task Discriminability* (mining hard negatives). Drawing upon GISTEmbed (Solatorio, 2024), we employ a frozen, powerful encoder to assess sample quality via *Dual-View Assessment*, leveraging the orthogonal perspectives of linguistic logic and visual intuition.

Guide Scoring with Visual Relaxation. We first quantify the plausibility of each sample $\langle q, pos \rangle$. For the textual score (s_{txt}), we verify semantic alignment via standard cosine similarity. For the visual score (s_{vis}), considering that generated images often capture *Visual Grouping Signals* (e.g., sibling resemblance) rather than precise hierarchical entailment, we adopt a visual relaxation strategy. Instead of strict matching, we accept the query q if it visually aligns with *any* constituent node of the candidate position pos:

$$\begin{aligned} s_{txt} &= \cos(\mathbf{e}_q^{txt}, \mathbf{e}_{pos}^{txt}) \\ s_{vis} &= \max_{n \in \{p,c,s\}} \cos(\mathbf{e}_q^{vis}, \mathbf{e}_n^{vis}) \end{aligned} \quad (4)$$

This *max* operation prevents false rejections when the image resembles a specific neighbor (e.g., a *Sibling* with shared shape) rather than the abstract *Parent*, effectively utilizing the grouping cues.

Consensus-Based Reweighting. We diagnose data quality based on cross-modal consensus and assign adaptive weights w. To overcome the limitations of single-modality filters (which risk over-cleaning due to *domain gaps* (Solatorio, 2024)), we design two complementary logic flows:

(1) Noise Handling (Mutual Rescue via Intersection): To address *Data Validity*, we identify noise only when *both* modalities reject the sample ($s_{txt} < \tau_n \wedge s_{vis} < \tau_n$). We apply *Soft Down-weighting* ($w^+ = \gamma_{noise} < 1$) rather than hard removal. This acts as a *Safety Net*, suppressing gradients from likely hallucinations while preserving valid long-tail concepts that are obscure in one modality but supported by the other.

(2) Hard Negative Mining (Complementary Union): To enhance *Task Discriminability*, we target the confusion caused by visual resemblance. If a negative sample is flagged as highly similar by *either* modality ($s_{txt} > \tau_h \vee s_{vis} > \tau_h$), we treat it as a *Hard Negative*. We *boost* its weight ($w^- = \gamma_{hard} > 1$) to impose a larger penalty, forcing the model to learn a rigorous decision boundary that distinguishes symmetric visual similarity (e.g., siblings) from asymmetric hierarchical entailment.

Final Objective. We employ a weighted pairwise contrastive margin loss. For a query q, its ground-truth position pos^+, and a set of sampled negatives $\mathcal{N}^-$, the loss is formulated as:

$$\mathcal{L} = w^+ D(q, pos^+)^2 + \sum_{n^- \in \mathcal{N}^-} w^- \left[m - D(q, n^-)\right]_+^2 \tag{5}$$

where $D(\cdot)$ denotes the cosine distance, m is the margin, and $[\cdot]_+$ denotes $\max(0, \cdot)$.

4 Experiment

4.1 Experimental Settings

Datasets and Metrics. We conduct experiments on three benchmarks following (Xu et al., 2023): SemEval-Food (Bordea et al., 2015) (concrete entities with distinct visual features), MeSH (Lipscomb, 2000) (specialized medical concepts), and WordNet-Verb (Jurgens and Pilehvar, 2016) (abstract actions challenging for visualization). Detailed statistics and splits are provided in Appendix A.1. Following standard protocols (Wang et al., 2022), we evaluate under the *all-rank* setting, reporting Macro Mean Rank (MR), Mean Reciprocal Rank (MRR), Recall@k, and Hit@k.

Baselines. We focus on *representation-based* TC methods, including TMN (Zhang et al., 2021), TaxoEnrich (Jiang et al., 2022), QEN (Wang et al., 2022), TaxoComplete (Arous et al., 2023), and CoSTC (Niu et al., 2024). Additionally, we adapt TE baselines TaxoExpan (Shen et al., 2020) and Arborist (Manzoor et al., 2020) to TC. We exclude *interaction-based* methods due to their prohibitive computational costs on large-scale taxonomies (Xu et al., 2025b). Note that we do not compare with multimodal TE methods (Zhu et al., 2023) due to the fundamental task mismatch. To clarify, the VITC (Text-Only) variant takes the *same* input as all text-only baselines: concept name plus original definition, with no visual pseudo-tokens. The detailed caption $\mathcal{L}_q$ is used *only* as the Stage 1 alignment target (Sec. 3.3) and never as a Stage 2 input. Any gain over VITC (Text-Only) thus reflects the contribution of visual injection, not richer text.

4.2 Main Results

Table 1 compares VITC against state-of-the-art baselines across three diverse datasets. First, VITC dominates on visually rich domains. On SemEval-Food, it achieves a remarkable 21.1% absolute gain in Hit@1 over the strongest baseline (CoSTC). This confirms that visual injection effectively bridges the "Sensory Gap", utilizing distinct visual cues to differentiate textually ambiguous concepts. Second, VITC exhibits strong robustness on abstract concepts. On WordNet-Verb and MeSH, VITC consistently outperforms text-only methods, notably doubling the Hit@1 on WordNet (14.6% $\rightarrow$ 31.3%). This validates that our generative injection strategy captures valid structural signals even for abstract nodes, while the MGAR module effectively prevents noise interference.

Impact of Multimodal Integration Architecture. Table 2 investigates different integration strategies. Note that to strictly isolate the impact of fusion architectures, we adapted the QEN baseline to use standard mean pooling (denoted as QEN (Mean Pooling)). While this simplification yields lower scores, it provides a uniform backbone for direct comparison with Late Fusion. (1) Validity of Visual Signals: Adding *Late Fusion* to this backbone improves performance, confirming visual cues indeed facilitate structural grounding. (2) Superiority of Deep Injection: Even without visual tokens, *VITC (Text-Only)* outperforms the multimodal *QEN + Late Fusion*, as our unified triplet template captures holistic structural interactions better than separate encoding (Niu et al., 2024). Crucially, Deep Injection delivers a substantial final leap (e.g., SemEval Hit@1 +16.2% over VITC Text-Only). This confirms that disjoint Late Fusion fractures the reasoning process, whereas injecting pseudo-tokens enables the text encoder to dynamically "consult" visual evidence during structural inference.

4.3 Ablation Studies

Is the Gain from Rich Text or Visual Injection? A natural concern is whether VITC's improvement stems from the visual modality itself or merely from reading a richer textual caption. To disentangle these factors, we feed the detailed caption to the text-only pipeline in two forms (Table 3). (1) Rich text alone causes semantic drift: replacing the strict definition with the scene-oriented caption drops Hit@1 from 43.9 to 33.8, as the model is distracted by descriptors (e.g., "wooden serving plate") deviating from strict is-a semantics. (2) The concise definition is indispensable: restoring it via Def+Caption recovers Hit@1 to 47.8, still substantially below our full model. (3) Visual tokens solve the information-overload problem: even with both definition and caption provided, Def+Caption lags

Table 1: Overall results on three datasets. The top section compares VITC with baselines, while the bottom section presents ablation studies for three proposed modules. Metrics are evaluated on total, leaf, and non-leaf nodes. ↓: lower is better. Best results are **bold**, and the best baseline results are underlined.

Datasets	Methods	Total								Leaf			Non-leaf		
		MR↓	MRR	R@1	R@5	R@10	H@1	H@5	H@10	MRR	H@5	R@10	MRR	H@5	R@10
SemEval-Food	TaxoExpan	371.291	0.286	5.7	13.3	18.0	11.5	26.4	34.5	0.477	30.1	35.6	0.130	8.0	3.6
	Arborist	256.491	0.290	13.0	18.0	21.0	26.4	34.5	38.5	0.466	39.0	38.5	0.146	12.0	6.7
	TMN	173.516	0.332	10.7	18.7	22.0	21.6	36.5	39.9	0.538	41.5	41.5	0.164	12.0	6.1
	TaxoEnrich	230.424	0.408	11.7	26.7	31.7	23.6	49.3	58.1	0.723	58.5	66.7	0.149	4.0	3.0
	QEN	336.554	0.439	21.9	30.9	35.0	45.9	58.8	64.9	0.732	64.2	68.9	0.209	32.0	9.1
	TaxoComplete	296.072	0.489	14.7	30.0	38.0	29.7	55.4	65.5	0.702	60.2	65.2	0.315	32.0	15.8
	CoSTC	61.471	0.658	18.7	43.0	54.3	39.0	73.4	80.4	0.825	74.5	78.0	0.529	68.0	**36.0**
	VITC	**19.875**	**0.692**	**28.6**	**49.8**	**58.5**	**60.1**	**82.4**	**87.1**	**0.902**	**84.6**	**88.1**	**0.531**	**72.0**	**36.0**
	-No Guided Filter	30.514	0.641	23.8	41.8	52.1	50.0	70.9	80.4	0.862	74.0	80.7	0.471	56.0	30.1
	-No Adaptive Boost	36.873	0.667	27.0	45.3	56.9	56.7	74.3	83.8	0.861	76.4	83.0	0.517	64.0	32.4
	-No Visual Alignment	30.610	0.632	25.1	41.8	52.1	52.7	72.9	79.7	0.875	77.2	82.9	0.446	56.0	28.4
MeSH	TaxoExpan	1029.344	0.233	2.7	6.2	12.2	6.0	12.7	23.9	0.381	16.3	24.3	0.137	5.0	4.3
	Arborist	843.199	0.337	5.0	13.6	21.8	11.0	25.8	37.4	0.437	26.7	30.6	0.271	23.8	16.0
	TMN	567.831	0.372	7.2	17.3	24.6	15.9	33.6	43.8	0.525	38.4	40.7	0.271	23.4	14.1
	TaxoEnrich	393.062	0.424	7.4	22.4	31.0	16.2	42.6	52.5	0.619	51.3	54.1	0.296	24.1	15.9
	QEN	451.253	0.438	7.5	21.3	30.8	17.1	43.1	55.9	0.611	51.1	51.8	0.332	26.1	17.9
	TaxoComplete	357.494	0.540	10.8	29.3	41.1	24.5	54.1	63.9	0.605	53.8	52.5	0.500	54.8	34.1
	CoSTC	109.081	0.600	11.0	34.6	47.5	24.9	61.5	72.6	0.741	63.5	66.7	0.512	57.4	35.7
	VITC	**45.798**	**0.687**	**19.4**	**44.2**	**57.2**	**44.2**	**75.1**	**83.2**	**0.835**	**78.8**	**77.3**	**0.597**	**67.1**	**44.8**
	-No Guided Filter	59.084	0.652	16.3	40.2	52.4	37.1	70.6	81.6	0.838	77.0	77.7	0.538	56.6	36.8
	-No Adaptive Boost	47.829	0.659	18.0	41.4	53.5	40.9	71.4	80.7	0.826	75.0	76.1	0.555	63.5	39.4
	-No Visual Alignment	43.967	0.633	17.3	38.9	51.5	39.4	68.8	78.8	0.782	71.5	71.2	0.540	63.2	39.3
WordNet-Verb	TaxoExpan	1752.271	0.215	4.1	11.4	15.1	6.1	17.1	22.5	0.354	20.5	26.7	0.057	3.1	1.7
	Arborist	1455.251	0.246	3.8	11.0	15.5	5.7	15.5	21.6	0.331	16.2	21.8	0.148	12.8	8.4
	TMN	1513.634	0.290	5.4	14.7	20.7	8.1	21.2	29.1	0.425	23.8	32.8	0.136	10.7	6.8
	TaxoEnrich	5462.075	0.179	3.9	9.0	12.3	5.8	13.6	18.4	0.313	16.8	22.6	0.025	0.5	0.4
	QEN	1730.755	0.404	9.1	23.3	31.0	13.9	34.0	43.9	0.568	38.6	48.4	0.224	15.3	11.8
	TaxoComplete	2661.488	0.407	9.0	22.2	30.9	13.6	31.7	40.8	0.487	32.7	41.3	0.315	27.6	19.1
	CoSTC	241.089	0.505	9.5	27.8	39.1	14.6	39.2	53.1	0.651	41.0	54.7	0.344	31.6	21.8
	VITC	**203.444**	**0.581**	**20.4**	**38.1**	**47.5**	**31.3**	**53.9**	**63.3**	**0.744**	**57.6**	**66.4**	**0.400**	**38.8**	**26.5**
	-No Guided Filter	161.915	0.552	15.3	35.3	44.5	23.5	49.9	60.3	0.710	53.1	62.7	0.377	36.7	24.3
	-No Adaptive Boost	167.566	0.567	16.4	36.0	46.9	25.2	50.0	61.2	0.716	53.4	63.7	0.400	36.2	26.0
	-No Visual Alignment	207.782	0.534	15.5	33.3	41.7	23.8	46.4	55.3	0.663	48.6	56.5	0.391	37.2	25.4

Table 2: Performance comparison of different multi-modal integration architectures. **QEN (Mean Pooling)** is a variant of QEN adapted with mean pooling to enable direct feature concatenation. **Late Fusion** concatenates visual features to this adapted backbone. **Ours** employs the proposed Deep Injection mechanism.

Datasets	Settings	MRR	H@1	H@5	R@5	R@10
SemEval-Food	QEN (Mean Pooling)	0.413	37.5	54.3	27.6	32.7
	+ Late Fusion	0.434	43.0	57.2	29.6	34.7
	VITC (Text-Only)	0.608	43.9	75.6	41.8	50.2
	Ours (Deep Injection)	**0.692**	**60.1**	**82.4**	**49.8**	**58.5**
MeSH	QEN (Mean Pooling)	0.343	19.2	37.9	18.4	25.3
	+ Late Fusion	0.408	21.3	42.6	20.3	27.6
	VITC (Text-Only)	0.592	31.8	65.4	34.5	46.3
	Ours (Deep Injection)	**0.687**	**44.2**	**75.1**	**44.2**	**57.2**
WordNet-Verb	QEN (Mean Pooling)	0.284	12.9	28.7	18.5	23.1
	+ Late Fusion	0.317	15.8	34.6	23.8	25.2
	VITC (Text-Only)	0.513	19.4	43.9	30.2	40.1
	Ours (Deep Injection)	**0.581**	**31.3**	**53.9**	**38.1**	**47.5**

VITC (Full) by 12.3% Hit@1. Concatenating long captions for three nodes $\langle p, c, s \rangle$ exceeds the 512-token limit, truncating critical definitions. VITC instead compresses the same rich semantics into k=4 pseudo-tokens, confirming the gain originates from *genuine visual grounding*, not longer text.

Table 3: **Textual Input** Ablation on SemEval-Food. *Caption* replaces the concise definition with the detailed image caption; *Def+Caption* concatenates both. VITC (Full) instead compresses visual semantics into k=4 pseudo-tokens.

Settings	MRR	H@1	H@5	R@5	R@10
VITC (Text-Only)	0.608	43.9	75.6	41.8	50.2
VITC (Caption)	0.539	33.8	64.2	34.1	39.6
VITC (Def+Caption)	0.647	47.8	72.9	42.4	52.7
VITC (Full, Ours)	**0.692**	**60.1**	**82.4**	**49.8**	**58.5**

Impact of Structure-Aware Visual Mapping. Table 4 validates the proxy tasks designed for visual projection. (1) Foundation of Visual Semantics: Task A acts as the primary information source. Removing it causes Stage 1 alignment metrics ($I \rightarrow C$, $I \rightarrow L$) to collapse to near-zero, dragging downstream performance down to the level of random initialization (*w/o Stage 1*). This confirms that reconstructing detailed captions is essential for encoding visual richness. (2) Structural Adaptation: Task C is crucial for bridging the gap to Stage 2 reasoning. Its removal leads to performance degra-

Table 4: Ablation studies of pre-training tasks in Stage 1. **Task A**: Richness, **Task B**: Concept, **Task C**: Structure. We report R@1 for Image-to-Concept (I→C) and R@5 for Image-to-LongCaption (I→L) in Stage 1, and ranking metrics for Stage 2.

Datasets	Settings	Stage 1		Stage 2		
		I→C	I→L	MRR	H@1	H@10
SemEval-Food	Ours	**15.5**	**61.5**	**0.692**	**60.1**	**87.1**
	w/o Task A	2.7	4.7	0.633	52.0	83.8
	w/o Task B	14.9	55.4	0.658	58.3	82.4
	w/o Task C	10.8	54.1	0.648	55.4	85.1
	w/o Stage 1	0.7	3.4	0.632	52.7	79.7
MeSH	Ours	**11.4**	**56.2**	**0.687**	**44.2**	**83.2**
	w/o Task A	0.2	1.3	0.646	40.1	79.6
	w/o Task B	8.4	51.3	0.673	43.8	81.2
	w/o Task C	7.0	49.2	0.671	40.2	82.6
	w/o Stage 1	0.1	0.5	0.633	39.4	78.8
WordNet-Verb	Ours	**5.7**	**59.5**	**0.581**	**31.3**	**63.3**
	w/o Task A	0.2	0.7	0.541	23.8	58.4
	w/o Task B	3.1	54.3	0.568	26.7	62.7
	w/o Task C	2.8	40.2	0.555	25.6	61.0
	w/o Stage 1	0.0	0.3	0.534	23.8	55.3

dation (e.g., -4.7% Hit@1 on Food), proving that pre-training within topological templates enables pseudo-tokens to function effectively in hierarchical contexts. (3) Semantic Intersection: Task B shows the smallest impact. We attribute this to the fact that the detailed captions in Task A often implicitly contain concept names, rendering Task B a semantic subset that serves primarily as a stabilizer rather than a new information source.

Impact of Adaptive Residual Fusion. Table 5 validates the decoupling strategy. (1) Calibrating Magnitude: *Naive Mean Pooling* fails due to density imbalance. Removing Visual Boost causes a sharp decline in Hit@1 (e.g., -5.4% on Food), confirming that scarce visual tokens require calibration to survive textual drowning. (2) Filtering Noise: Removing Adaptive Gating degrades performance, proving unconditional boosting amplifies noise, particularly in abstract domains. (3) Explicit Fusion Necessity: *w/o Image Pooling* outperforms the Pure Text baseline (Table 2) due to implicit attention and MGAR filtering. However, *Ours* yields a distinct gain, confirming that while "consulting" pseudo-tokens helps, explicitly fusing visual residuals is critical for maximizing discriminability.

Impact of MGAR Strategy (Logic). Table 6 validates the cross-modal consensus logic. (1) Hard Negative Mining (Union): Compared to text-only mining, the *MM Hard Neg* setting consistently improves performance (e.g., MeSH Hit@1 +2.5%), confirming that visual similarity helps identify confusing siblings that text misses (Complementary Mining). (2) Noise Filtering (Intersection): A critical observation is the "Over-cleaning Paradox": applying a *Text Noise Filter* actually *degrades* performance across all three datasets compared to the previous step. This proves that single-modality filtering erroneously discards valid long-tail concepts. However, our MM Noise Filter not only recovers this loss but achieves the best overall performance, validating the Mutual Rescue mechanism: preserving samples unless *both* modalities reject them acts as a crucial safety net.

Impact of Weighting Actions (Implementation). Table 7 analyzes the processing actions. For Noise (Part 1), *Soft Down-weighting* outperforms Hard

Table 5: Ablation studies of the fusion mechanism in Stage 2. We investigate the effectiveness of the proposed **Adaptive Residual Fusion** by removing specific components. *Naive Mean Pooling* denotes the standard pooling strategy without separation. *w/o Image Pooling* represents the setting where image tokens are input to the encoder for attention interaction but excluded from the final pooling aggregation.

Datasets	Settings	MRR	H@1	H@5	R@5	R@10
SemEval-Food	Ours	**0.692**	**60.1**	**82.4**	**49.8**	**58.5**
	w/o Visual Boost	0.665	54.7	79.7	46.0	55.3
	w/o Adaptive Gating	0.637	54.7	75.0	45.0	53.1
	Naive Mean Pooling	0.667	56.7	74.3	45.3	56.9
	w/o Image Pooling	0.643	54.0	74.6	45.1	54.0
MeSH	Ours	**0.687**	**44.2**	**75.1**	**44.2**	**57.2**
	w/o Visual Boost	0.661	41.8	73.5	42.7	54.9
	w/o Adaptive Gating	0.663	41.2	72.8	42.0	54.4
	Naive Mean Pooling	0.659	40.9	71.4	41.4	53.5
	w/o Image Pooling	0.652	36.7	71.0	39.9	52.8
WordNet-Verb	Ours	**0.581**	**31.3**	**53.9**	**38.1**	**47.5**
	w/o Visual Boost	0.563	28.6	52.9	37.1	46.1
	w/o Adaptive Gating	0.566	26.9	52.2	36.9	46.6
	Naive Mean Pooling	0.567	25.2	50.0	36.0	46.9
	w/o Image Pooling	0.558	24.4	51.9	36.2	45.5

Table 6: Step-by-step ablation studies of the MGAR module. We progressively add components to validate the effectiveness of the proposed **Multi-modal Hard Mining (Union)** and **Multi-modal Noise Filtering (Intersection)**.

Datasets	Settings	MRR	H@1	H@5	R@5	R@10
SemEval-Food	Vanilla (Rand. Neg.)	0.641	50.0	70.9	41.8	52.1
	+ Text Hard Neg	0.676	55.4	77.0	46.8	57.6
	+ MM Hard Neg (Union)	0.687	56.1	78.4	46.9	58.5
	+ Text Noise Filter	0.644	54.1	77.7	45.3	53.4
	+ MM Noise Filter (Ours)	**0.692**	**60.1**	**82.4**	**49.8**	**58.5**
MeSH	Vanilla (Rand. Neg.)	0.652	37.1	70.6	40.2	52.4
	+ Text Hard Neg	0.670	40.4	71.9	42.1	55.2
	+ MM Hard Neg (Union)	0.675	42.9	72.9	42.1	56.1
	+ Text Noise Filter	0.662	41.2	71.2	41.6	54.4
	+ MM Noise Filter (Ours)	**0.687**	**44.2**	**75.1**	**44.2**	**57.2**
WordNet-Verb	Vanilla (Rand. Neg.)	0.552	23.5	59.9	38.1	47.5
	+ Text Hard Neg	0.567	27.0	53.4	37.3	46.8
	+ MM Hard Neg (Union)	0.577	28.7	53.5	38.0	47.5
	+ Text Noise Filter	0.559	26.7	52.1	36.5	45.6
	+ MM Noise Filter (Ours)	**0.581**	**31.3**	**53.9**	**38.1**	**47.5**

Table 7: Ablation study on weighting strategies for noise and hard negative handling on SemEval-Food. We report the impact of different weights (w) for **Noise Handling** (Part 1) and **Hard Negative Handling** (Part 2).

Action Logic	Weight (w)	Overall		Leaf (L)		Non-Leaf (NL)	
		MRR	H@5	MRR	H@1	MRR	H@1
Part 1: Noise Handling *(Fix Hard Neg $w = 2.0$)*							
Hard Removal	0.0	0.687	79.7	0.900	56.1	0.525	36.0
Soft Down-weight (Ours)	**0.1**	**0.692**	**82.4**	**0.902**	**64.2**	**0.531**	**40.0**
Part 2: Hard Negative Handling *(Fix Noise $w = 0.1$)*							
False Neg Filter	0.0	0.663	70.3	0.786	48.8	**0.569**	**48.0**
Ignore (Standard)	1.0	0.651	76.4	0.881	56.1	0.474	20.0
Soft Up-weight (Ours)	**2.0**	**0.692**	**82.4**	**0.902**	**64.2**	0.531	40.0

Table 8: Impact of generator and captioner choices on SemEval-Food. *I→L* reports Stage 1 alignment quality (R@5); *H@1* and *MRR* report Stage 2 performance.

Generator	Captioner	I→L	H@1	MRR
DALL-E 3	Internal	61.5	**60.1**	**0.692**
SDXL	LLaVA-8B	57.4	59.5	0.675
SDXL	BLIP-2	33.8	58.7	0.674
SDXL	BLIP	**62.2**	52.0	0.663
Text-Only (No Visual)		–	43.9	0.608

Removal, indicating that flagged samples still retain partial valid signals (e.g., node names) useful for weak supervision. For Hard Negatives (Part 2), discarding them (w=0.0) degrades performance, particularly on Leaf nodes. Conversely, *Up-weighting* boosts discriminability, confirming that visually similar siblings are not noise, but critical boundary-defining samples that the model must actively learn to distinguish.

Impact of Generator and Captioner Choice. To address the concern of dependency on proprietary DALL-E 3, we replace the generation pipeline with fully open-source alternatives: Stable Diffusion XL (SDXL) paired with various captioners (Table 8). (1) *Open-source feasibility*: the SDXL+LLaVA pipeline reaches 59.5 Hit@1, nearly matching DALL-E 3 (60.1) and far exceeding the text-only baseline (43.9), confirming VITC is not bound to proprietary generators and remains reproducible with open tools. (2) *Captioner quality dominates*: replacing LLaVA with lightweight captioners sharply degrades Hit@1, even though BLIP attains a higher Stage 1 score (62.2). This counterintuitive gap reveals that shallow captions inflate alignment metrics yet fail to capture the fine-grained visual details required by Richness Distillation (Task A), validating its role as the primary information source in Stage 1. We note that prompt and random seed variations constitute minor perturbations compared to switching major pipeline components (generator/captioner), and our results remain stable across such variations.

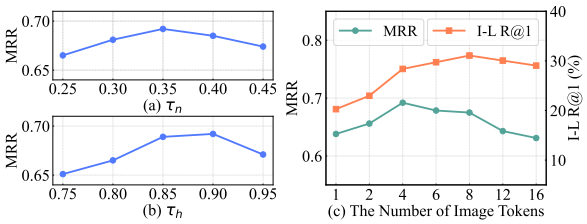

Figure 3: Parameter Sensitivity Analysis on SemEval-Food. (a) Noise filtering threshold τ_n. (b) Hard negative mining threshold τ_h. (c) Impact of visual token number k, comparing alignment accuracy (I-L R@1) in Stage 1 vs. downstream ranking (MRR) in Stage 2.

Parameter Study. Figure 3 investigates the impact of key hyperparameters. **(1) Thresholds (τ_n, τ_h):** Both exhibit inverted U-shaped trends. Low τ_n (< 0.3) fails to filter hallucinations, while high values (> 0.4) over-clean valid concepts. For τ_h, a moderate 0.9 is optimal; lower values confuse distinct concepts as hard negatives, while higher values miss subtle sibling cues. **(2) Token Number (k): A Trade-off.** Figure 3(c) illustrates the *Density-Selectivity Dilemma*. While increasing k from 1 to 8 improves Stage 1 reconstruction quality (I-L R@1 rises), it harms downstream TC performance (MRR drops after $k = 4$). This indicates that although more tokens encode richer visual details, they also introduce redundancy and numerically overwhelm the textual context in the fusion stage. Thus, $k = 4$ strikes the optimal balance between information richness and token density.

5 Conclusion

We proposed VITC to bridge the "Sensory Gap" in Taxonomy Completion by injecting synthesized images as pseudo-tokens. Adaptive Residual Fusion and the consensus-driven MGAR strategy address the resulting modality imbalance and data quality challenges. Experiments on three datasets yield absolute Hit@1 gains of 21.1%, 19.3%, and 16.7% over the strongest baselines. Ablations confirm the gain stems from genuine visual grounding rather than richer text, and that deep injection outperforms late fusion by enabling the encoder to consult visual cues during structural reasoning. The framework generalizes to open-source pipelines, offering a reproducible pathway for multimodal knowledge engineering where visual data is unavailable.

Limitations

While VITC demonstrates the effectiveness of generative visual injection, we acknowledge several limitations that suggest avenues for future improvement:

(1) **Dependency on Generative Models.** Our framework relies on an external generator to synthesize visual signals. While VITC is compatible with both proprietary (DALL-E 3) and fully open-source (SDXL) pipelines (Sec. 4.3), the final performance remains contingent on the generator's capacity to produce faithful images. Moreover, while offline synthesis is a one-time overhead, the computational resources required for large-scale preparation are notably higher than those of retrieval-based methods. For polysemous definitions, generators may follow the wrong sense and amplify errors (Appendix A.8).

(2) **Non-Visualizable Concepts.** While robust on abstract verbs (WordNet) and medical terms (MeSH), there remains a fundamental limit to visual grounding. Extremely abstract concepts, such as philosophical terms (e.g., *"Existentialism"*) or grammatical function words, often lack meaningful visual representations. Such concepts create an inherent gap between linguistic richness and visual grounding that remains difficult to bridge, potentially limiting the gain from visual injection. In such cases, the model primarily falls back to textual semantics via the gating mechanism.

(3) **Imperceptible and Hierarchical Traits.** Case studies (Appendix A.8) reveal two failure modes: (i) concepts defined by internal composition rather than appearance (e.g., *Skim Milk*, defined by fat content), and (ii) parent-child pairs that are visually near-identical (e.g., *Sparkling Wine* vs. *Champagne*), where vision excels at distinguishing siblings but struggles with hierarchical inclusion. Tightly integrating structural priors with visual evidence to capture these fine-grained distinctions remains an open direction.

Acknowledgments

We sincerely thank anonymous reviewers for their valuable comments. We also thank Xinzhu Sun for her insightful suggestions and inspiring ideas on the aesthetic design of figures and tables in this paper. This research is supported by the National Natural Science Foundation of China (No. 72342017, 62572260). Computation is supported by the Supercomputing Center of Nankai University (NKSC).

We utilized AI assistants solely for the purpose of refining the writing style and checking grammatical errors in the manuscript.

References

Lorenzo Agnolucci, Alberto Baldrati, Alberto Del Bimbo, and Marco Bertini. 2025. isearle: Improving textual inversion for zero-shot composed image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(11):10801–10817.

Ines Arous, Ljiljana Dolamic, and Philippe Cudré-Mauroux. 2023. Taxocomplete: Self-supervised taxonomy completion leveraging position-enhanced semantic matching. In *WWW*, pages 2509–2518.

Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. 2023. Zero-shot composed image retrieval with textual inversion. In *ICCV*, pages 15292–15301.

Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. 2022. Effective conditioned and composed image retrieval combining clip-based features. In *CVPR*, pages 21434–21442.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, and 1 others. 2023. Improving image generation with better captions. *Computer Science.*, 2(3):8.

Georgeta Bordea, Paul Buitelaar, Stefano Faralli, and Roberto Navigli. 2015. Semeval-2015 task 17: Taxonomy extraction evaluation (texeval). In *SemEval@NAACL-HLT*, pages 902–910.

Niv Cohen, Rinon Gal, Eli A. Meirom, Gal Chechik, and Yuval Atzmon. 2022. This is my unicorn, fluffy: Personalizing frozen vision-language representations. In *ECCV*, pages 558–577.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-Or. 2023. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *EMNLP*, pages 6894–6910.

Minhao Jiang, Xiangchen Song, Jieyu Zhang, and Jiawei Han. 2022. Taxoenrich: Self-supervised taxonomy completion via structure-semantic representations. In *WWW*, pages 925–934.

David Jurgens and Mohammad Taher Pilehvar. 2016. Semeval-2016 task 14: Semantic taxonomy enrichment. In *SemEval-2016*, pages 1092–1102.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. BLIP: bootstrapping language-image pre-training for unified vision-language understanding

and generation. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.

Carolyn E Lipscomb. 2000. Medical subject headings (mesh). *Bulletin of the Medical Library Association*, page 265.

Zichen Liu, Hongyuan Xu, Yanlong Wen, Ning Jiang, Haiying Wu, and Xiaojie Yuan. 2021. TEMP: taxonomy expansion with dynamic margin loss through taxonomy-paths. In *EMNLP*, pages 3854–3863.

Emaad Manzoor, Rui Li, Dhananjay Shrouty, and Jure Leskovec. 2020. Expanding taxonomies with implicit edge semantics. In *WWW*, pages 2044–2054.

Yuhang Niu, Hongyuan Xu, Ciyi Liu, Yanlong Wen, and Xiaojie Yuan. 2024. Contrastive representation learning for self-supervised taxonomy completion. In *IJCAI*, pages 6442–6450.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Trans. Knowl. Data Eng.*, 36(7):3580–3599.

Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2021. Contrastive learning with hard negative samples. In *ICLR*.

Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. 2023. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *CVPR*, pages 19305–19314.

Jiaming Shen, Zhihong Shen, Chenyan Xiong, Chi Wang, Kuansan Wang, and Jiawei Han. 2020. Taxoexpan: Self-supervised taxonomy expansion with position-enhanced graph neural network. In *WWW*, pages 486–497.

Jingchuan Shi, Hang Dong, Jiaoyan Chen, Zhe Wu, and Ian Horrocks. 2024. Taxonomy completion via implicit concept insertion. In *WWW*, pages 2159–2169.

Aivin V. Solatorio. 2024. Gistembed: Guided in-sample selection of training negatives for text embedding fine-tuning. *CoRR*, abs/2402.16829.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. In *NeurIPS*.

Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. 2015. Training very deep networks. In *NeurIPS*, pages 2377–2385.

Qwen Team. 2025. Qwen3 technical report. *CoRR*, abs/2505.09388.

Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. 2023. P+: extended textual conditioning in text-to-image generation. *CoRR*, abs/2303.09522.

Suyuchen Wang, Ruihui Zhao, Yefeng Zheng, and Bang Liu. 2022. Qen: Applicable taxonomy completion via evaluating full taxonomic relations. In *WWW*, pages 1008–1017.

Haokun Wen, Xuemeng Song, Xiaolin Chen, Yinwei Wei, Liqiang Nie, and Tat-Seng Chua. 2024. Simple but effective raw-data level multimodal fusion for composed image retrieval. In *SIGIR*, pages 229–239. ACM.

Hongyuan Xu, Ciyi Liu, Yuhang Niu, Yunong Chen, Xiangrui Cai, Yanlong Wen, and Xiaojie Yuan. 2023. Tacoprompt: A collaborative multi-task prompt learning method for self-supervised taxonomy completion. In *EMNLP*, pages 15804–15817.

Hongyuan Xu, Yuhang Niu, Ciyi Liu, Yanlong Wen, and Xiaojie Yuan. 2025a. Taxopro: A plug-in lora-based cross-domain method for low-resource taxonomy completion. *Trans. Assoc. Comput. Linguistics*, 13:557–576.

Hongyuan Xu, Yuhang Niu, Yanlong Wen, and Xiaojie Yuan. 2025b. Compress and mix: Advancing efficient taxonomy completion with large language models. In *WWW*, pages 4239–4249. ACM.

Wei Xue, Yongliang Shen, Wenqi Ren, Jietian Guo, Shiliang Pu, and Weiming Lu. 2024. Insert or attach: Taxonomy completion via box embedding. In *ACL*, pages 3851–3863.

Qingkai Zeng, Jinfeng Lin, Wenhao Yu, Jane Cleland-Huang, and Meng Jiang. 2021. Enhancing taxonomy completion with concept generation via fusing relational representations. In *SIGKDD*, pages 2104–2113.

Jieyu Zhang, Xiangchen Song, Ying Zeng, Jiaze Chen, Jiaming Shen, Yuning Mao, and Lei Li. 2021. Taxonomy completion via triplet matching network. In *AAAI*, pages 4662–4670.

Yuchen Zhang, Amr Ahmed, Vanja Josifovski, and Alexander J. Smola. 2014. Taxonomy discovery for personalized recommendation. In *WSDM*, pages 243–252. ACM.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Conditional prompt learning for vision-language models. In *CVPR*.

Tinghui Zhu, Jingping Liu, Jiaqing Liang, Haiyun Jiang, Yanghua Xiao, Zongyu Wang, Rui Xie, and Yunsen Xian. 2023. Towards visual taxonomy expansion. In *ACM MM*, pages 6481–6490.

A Appendix

A.1 Dataset Details

We conduct experiments on three public datasets covering diverse domains and abstraction levels. Statistical details are summarized in Table 9.

- **SemEval-Food** (Bordea et al., 2015): Originating from SemEval-2015 Task 17, this dataset constructs a hierarchical taxonomy specifically for the food domain, representing concrete entities with distinct visual features.

- **MeSH (Medical Subject Headings)** (Lipscomb, 2000): A subgraph extracted from the comprehensive biomedical indexing system. It represents a specialized clinical taxonomy characterized by rigorous scientific definitions.

- **WordNet-Verb** (Jurgens and Pilehvar, 2016): Derived from SemEval-2016 Task 14 (based on WordNet 3.0), this dataset organizes verbs into a hypernym hierarchy. It presents a unique challenge due to the high abstraction and polysemy of action-oriented concepts.

Data Partitioning. Following (Wang et al., 2022; Xu et al., 2023), we partition the node set $\mathcal{N}$ into disjoint subsets: $\mathcal{N}_{\text{train}}$, $\mathcal{N}_{\text{validation}}$, and $\mathcal{N}_{\text{test}}$. For **SemEval-Food** and **MeSH**, we adopt a percentage-based split: 10% of nodes are allocated for validation and 10% for testing, with the remaining 80% serving as the training set. For **WordNet-Verb**, given its scale, we randomly sample fixed sets of 1,000 nodes each for validation and testing, retaining the rest for training.

Table 9: Detailed statistics of the three benchmark datasets. $|\mathcal{N}|$ and $|\mathcal{E}|$ denote the total count of nodes and edges. #depth indicates the maximum hierarchical depth, and #avg.tokens represents the average length of textual definitions. #candidates refers to the total number of candidate positions considered during evaluation.

Dataset	$\lvert\mathcal{N}\rvert/\lvert\mathcal{N}_{\text{train}}\rvert$	$\lvert\mathcal{E}\rvert$	#depth	#avg.tokens	#candidates
SemEval-Food	1,486 / 1,190	1,533	8	34.6	7,313
MeSH	9,710 / 8,072	10,498	10	62.6	42,970
WordNet-Verb	13,936 / 11,936	13,407	12	26.4	51,159

A.2 Baseline Descriptions

We compare our framework against the following state-of-the-art representation-based taxonomy completion methods:

- **TMN** (Zhang et al., 2021): A channel-wise matching framework that enhances concept representation by decomposing the main task into auxiliary sub-objectives, specifically matching the query to potential parents and children to the query.

- **TaxoEnrich** (Jiang et al., 2022): This method focuses on capturing the local topological context. It employs a query-aware sibling aggregator to refine position representations using taxonomy-contextualized embeddings.

- **QEN** (Wang et al., 2022): Leveraging pre-trained language models, QEN generates semantic representations with a specific focus on mitigating noise from "pseudo-leaves" via a sibling-relation objective.

- **TaxoComplete** (Arous et al., 2023): A self-supervised framework that combines a bi-encoder architecture for semantic matching with a direction-aware graph propagation mechanism to generate position-enhanced node embeddings.

- **CoSTC** (Niu et al., 2024): Adopts a contrastive learning paradigm to extract taxonomic relations. It constructs two distinct contrastive views and employs a hard negative sampling strategy to improve discriminative capability.

A.3 Implementation Details

We employ `MPNet-base` (Song et al., 2020) as the text encoder and the visual branch of `blip-itm-large-coco` (Li et al., 2022) as the frozen vision backbone. For the Guide Scoring in the MGAR strategy (Sec. 3.5), we extract the text embeddings e^{txt} using the frozen `Qwen3-Embedding-0.6B` (Team, 2025). For the visual embeddings e^{vis}, we reuse the frozen representations extracted by the aforementioned `blip-itm-large-coco` vision backbone prior to the mapping network. These guide models remain strictly frozen during the entire training process to provide stable reference scores. The mapping network is a 2-layer MLP (increasing from 256 to 512, and finally 768, or as configured) with LayerNorm and GELU. Training is conducted on an NVIDIA A800 using AdamW: (1) **Stage 1:** We optimize only the mapping network for 60 epochs (BS=128, LR=$1e^{-4}$). The anti-shortcut dropout rates are set to $\rho_{def} = 0.3$ and $\rho_{token} = 0.1$. (2) **Stage 2:** We fine-tune the text encoder and mapping network for 20 epochs (BS=256, LR=$5e^{-5}$). To ensure **Robust Initialization** of the Adaptive Fusion module, the gate bias $\mathbf{b}_g$ is initialized to -2.0, starting with a near-zero visual residual. For **MGAR**, we set the consensus

threshold $\tau_n = 0.35, \tau_h = 0.9$, margin $m = 0.5$, and adaptive weights $\gamma_{noise} = 0.1, \gamma_{hard} = 2.0$. Following (Wang et al., 2022), we add siblings to the candidate position:$\langle p, c, s \rangle$ by randomly selecting a child of p.

A.4 Visual Imputation Details

We generate representative images (1024×1024) for all taxonomy nodes using DALL-E 3. Note that this is a one-time, offline pre-processing step that incurs no computational overhead during the inference phase. To ensure coverage and safety, we employ a hierarchical three-step strategy:

1. Standard Generation. We primarily prompt the model with: *"Please generate the image of concept "[**Term**]". Its definition is "[**Definition**]"."*. Including the definition explicitly resolves polysemy.

2. Safety Fallback. Upon content policy violations (e.g., sensitive medical imagery), we retry with a "Scientific Illustration" style: *"An abstract, educational diagram representing: "[**Definition**]" (Concept: "[**Term**]"). Minimalist line art, safe for work."*. This abstract style reduces safety triggers while preserving structural semantics.

3. Synthetic Placeholder. For nodes where generation remains impossible (e.g., due to persistent safety/copyright restrictions) or artificial structural nodes (i.e., *pseudo-roots* and *pseudo-leaves*), we programmatically render the term's name as white text on a black background. This ensures the vision encoder extracts non-zero text-pattern features.

Notably, our empirical statistics observe that the Safety Fallback (Step 2) successfully resolved content policy restrictions for the vast majority of taxonomy concepts. Consequently, the Synthetic Placeholder (Step 3) was triggered primarily for structural auxiliary nodes (i.e., pseudo-roots and pseudo-leaves). With less than 1% of actual concepts reverting to placeholders, this guarantees that the system is predominantly grounded in rich, generative visual semantics, rendering the impact of OCR-based features statistically negligible.

Caption Alignment. For Stage 1 training, we use the `revised_prompt` returned by the DALL-E 3 API as the ground-truth caption, as it reflects the system's actual output more accurately than the input definition.

A.5 Qualitative examples

We visualize representative generated images across three datasets to analyze the quality of visual injection in Figure 4. Columns 1-4 (Good Matching) demonstrate that for concrete entities (e.g., "Beef Burrito" in SemEval) and distinct actions (e.g., "Judge" in WordNet), the generated images provide rich, discriminative visual cues that complement textual descriptions, validating our "Visual Imputation" strategy. Columns 5-6 (Abstract/Hard Cases) highlight the limitations of visualization for abstract concepts. For terms like "Mental dysfunction" (MeSH) or "Aggrieve" (WordNet), the images rely on metaphors or artistic abstractions rather than concrete evidence. This observation justifies our design of the Adaptive Residual Fusion mechanism, which acts as a "safety valve" to filter out such visual noise and down-weight ambiguous signals for non-visualizable concepts.

Beyond the concrete examples in Figure 4, we emphasize that synthesized images for abstract or high-level concepts should not be interpreted as literal visual ground truth. Rather, they often exhibit recurring visual metaphors or domain-level regularities, which can be captured by a separate frozen vision encoder and used as coarse discriminative cues for taxonomy completion. In other words, the generator provides a structured visual prior, while the downstream encoder exploits the resulting regularities to support structural matching. This perspective helps explain why VITC still improves performance on abstract or non-leaf nodes, despite the lack of directly visualizable physical attributes. Nevertheless, these signals are not always reliable: for highly non-visualizable concepts, the generated imagery may become overly artistic, ambiguous, or semantically unstable. This is consistent with the hard cases in Figure 4 and the failure analyses in Appendix A.8, and motivates our use of adaptive fusion to reduce the influence of uncertain visual evidence. For transparency and reproducibility, we release the full set of synthesized images in our public repository, so that interested readers can inspect a broader range of abstract and concrete concepts beyond the representative examples shown in Figure 4.

A.6 Computational Cost and Deployment

A common concern for multimodal frameworks is the computational overhead introduced by the visual pipeline. We clarify that VITC's heavy computation is strictly **offline** and **one-time**, while the online inference cost is negligible. Table 10 summarizes the end-to-end cost of each stage on a single NVIDIA A800 GPU.

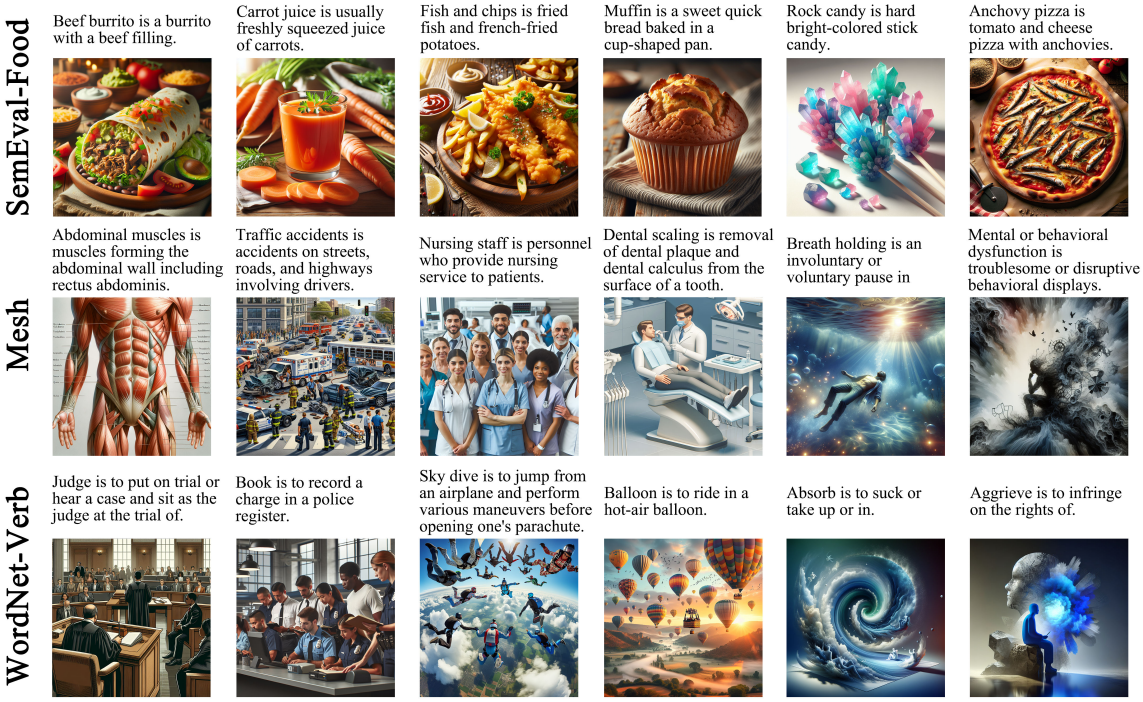

Figure 4: Illustration of generated visual images with concepts' definitions across three datasets.

Table 10: Cost breakdown of VITC. All heavy computations are performed offline; online inference adds only k=4 visual pseudo-tokens per node, yielding negligible latency.

Stage	Operation	Time	Mode
Pre-process	Image Gen. (SDXL)	∼2.0 s/concept	Offline
Pre-process	Captioning (LLaVA)	∼10 s/concept	Offline
Stage 1	Visual Mapping	<1 hr (total)	Offline
Stage 2	TC Fine-tuning	<30 min/epoch	Offline
Inference	Taxonomy Completion	<10 ms/query	Online

Offline Pre-processing. Image synthesis and captioning are executed once per concept and stored as cached features. All SDXL generation and LLaVA captioning runs use `batch size = 1` with FP16, fitting comfortably within the **low-teens GB range** on a single 24 GB consumer GPU (e.g., RTX 3090/4090).

Online Deployment. During inference, VITC only processes k=4 additional visual pseudo-tokens per node beyond the standard text input. The resulting latency increase over a plain BERT-base text encoder is negligible in practice, ensuring that visual injection does not harm deployment efficiency. Consequently, VITC can be trained and deployed entirely on a single consumer GPU, rendering it accessible for typical research settings.

A.7 Bias Audit of Generated Visual Data

Since generative models are known to encode societal biases, we conducted a systematic audit of the synthesized images and their associated captions used in VITC. We examine potential bias at two levels: visual content and textual descriptors.

Visual Content: Stereotypes vs. Discriminative Features. We inspected sensitive concepts involving human professions (e.g., *Nurse*, *Judge*, *Police Officer*). In contrast to well-known biases such as "Nurse = Female", the generated images exhibit **notable demographic diversity** in gender and ethnicity, consistent with the safety alignments of modern generators (DALL-E 3 / SDXL). We further distinguish *stereotypes* from *discriminative features*. Visual cues such as a *stethoscope* for a doctor or a *gavel* for a judge are **functional archetypes**, not demographic biases. They serve as essential features for VITC to separate e.g. *Doctor* from *Lawyer*. Our inspection confirms that these functional attributes are consistently present across samples, while demographic traits vary, allowing the model to learn the former without overfitting to the latter.

Textual Descriptors: Spurious Correlations. We also observed that a small fraction of captions

contain safety-related descriptors (e.g., "diverse group", "Caucasian"). We verified that these noise words do not alter hierarchy predictions, as Stage 1 (training the mapping network) implicitly filters out non-structural text. Nevertheless, such descriptors could in principle be stripped via a simple keyword filter before training, offering a straightforward mitigation path for safety-sensitive deployments.

Summary. The audit confirms that VITC predominantly leverages functional, task-relevant visual cues rather than demographic stereotypes. Combined with the MGAR noise-filtering mechanism (Sec. 3.5), the risk of propagating harmful bias into downstream taxonomy structures is effectively mitigated.

A.8 Qualitative Case Studies

To provide intuition beyond aggregate metrics, we present representative cases where VITC succeeds or fails relative to text-only baselines in Figure 5. Each case is structured as *Error / Visual Signal / Outcome / Insight*.

A.8.1 Success Cases: Bridging the Sensory Gap

Case 1 — Visual Texture Differentiation (*Emmenthal* vs. *Gruyere*). Error: Text-only baselines (e.g., TaxoComplete) incorrectly predict *Emmenthal* as a parent/child of *Gruyere*. While their definitions technically differ by a single modifier ("large" vs. "small" holes), they share extensive lexical overlap ("is Swiss cheese with..."). Text encoders struggle to prioritize this subtle single-word distinction over the overwhelming shared context, falling into a "lexical overlap trap" and inferring a false hierarchy. **Visual Signal:** The generated images faithfully translate this subtle textual difference into explicit, salient visual features. *Emmenthal* is depicted with characteristic prominent, large holes, contrasting clearly with the distinctly smaller pores of *Gruyere*. **Outcome:** By grounding these concepts visually, VITC overrides the deceptive textual template similarity and correctly identifies them as mutually exclusive *Siblings* under the shared parent *Swiss Cheese*. **Insight:** Visual injection acts as a magnifying glass for subtle textual differences. It transforms single-word modifiers—which are easily drowned out in text sequence matching—into dominant, highly discriminative visual textures.

Case 2 — Object-Level Discrimination (*Pepperoni Pizza* vs. *Anchovy Pizza*). Error: Baselines conflate the two concepts due to structurally identical definitions ("tomato and cheese pizza with *[topping]*"), a classic *lexical overlap trap*. **Visual Signal:** The image for *Pepperoni Pizza* features *evenly distributed red circular slices*, visually distinct from the dark, elongated strips of anchovies. **Outcome:** VITC overrides the textual template similarity and places *Pepperoni Pizza* at its correct leaf position. **Insight:** Object-level visual cues provide a discriminative signal orthogonal to repetitive textual templates.

A.8.2 Failure Cases: Limitations of Visual Injection

Case 3 — Noise Amplification via Polysemy (*Bourbon*). Error: The query *Bourbon* (intended as whiskey) is incorrectly described in the dataset as "a reactionary politician from the US South". **Visual Signal:** VITC's image generator faithfully follows this wrong definition, producing "a politician with an American flag". The strong but misleading visual signal overwhelms latent taxonomic cues. **Outcome:** The model drifts from *Whiskey* toward *K-ration* (military/US history). **Insight:** Visual injection is a *double-edged sword*: when the textual definition is itself noisy, generated images may amplify rather than correct the error. This motivates the cross-modal consensus design of MGAR, though cases where text and vision jointly err remain an open challenge.

Case 4 — Hierarchical Collapse (*Sparkling Wine* vs. *Champagne*). Error: The goal is to insert *Sparkling Wine* as a parent of *Champagne*. VITC instead places it as a child/sibling. **Visual Signal:** The generic parent (*Sparkling Wine*) and the specific child (*Champagne*) are visually near-identical: both feature "champagne flutes, bubbles, and celebratory vibes". **Outcome:** The model struggles to distinguish the generic class from its specific instance based on vision alone, leading to hierarchical collapse. **Insight:** Visual modalities excel at *"Difference" (siblings)* but struggle with *"Inclusion" (parent-child)*. This validates Adaptive Residual Fusion, which learns to down-weight the visual contribution (via the gate g) when visual distinctiveness is low, preventing vision from dominating structural reasoning.

Case 5 — Visually Imperceptible Attributes (*Skim Milk*). Error: *Skim Milk* is defined by its

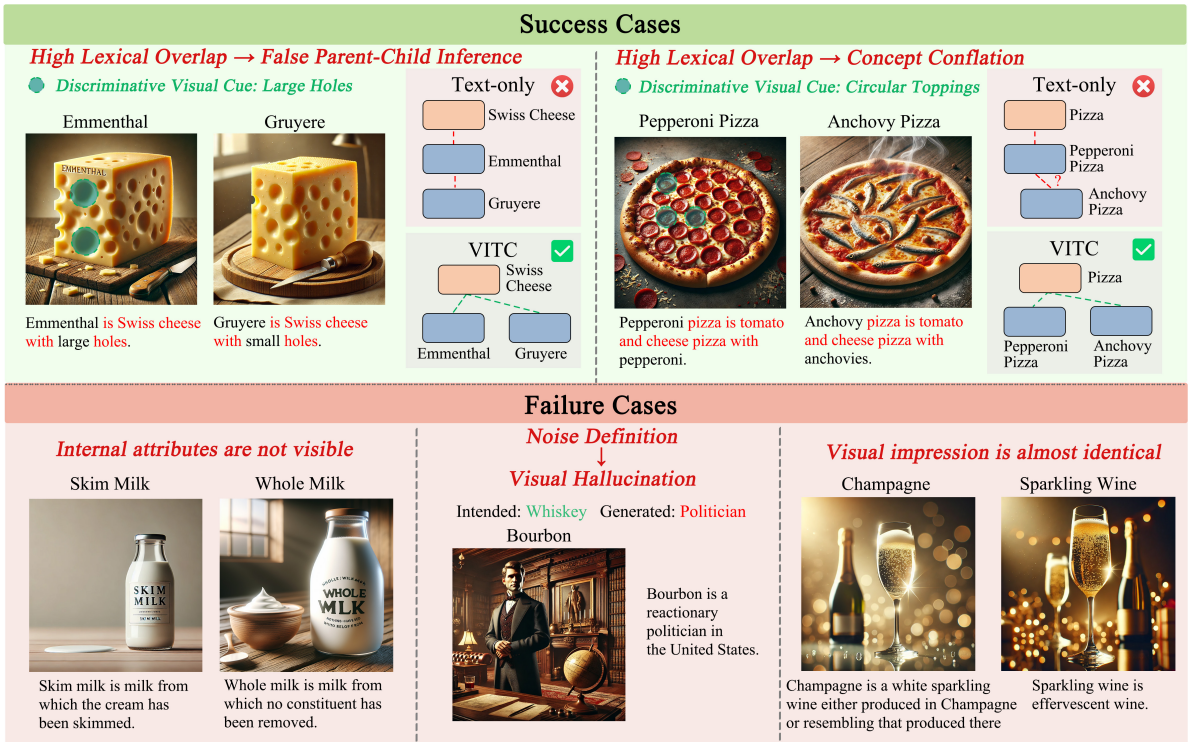

Figure 5: Case studies of VITC overcoming the Sensory Gap and its limitations.

fat content ("milk from which the cream has been skimmed"), a chemical attribute invisible to the eye. **Visual Signal:** The generated image is "pure white liquid in a clear glass", visually identical to *Whole Milk* or generic *Dairy*. The defining attribute (lack of fat) is lost in translation. **Outcome:** VITC is confused by the high visual similarity and fails to separate the specific subtype from its generic parent. **Insight:** Visual injection is less effective for concepts defined by internal composition rather than external appearance. This motivates the gating mechanism g of Adaptive Residual Fusion, which down-weights non-discriminative visual signals to let precise textual definitions dominate.