GRAN-TED: Generating Robust, Aligned, and Nuanced Text Embedding for Diffusion Models

Bozhou Li[1,2*†] Sihan Yang[1,3*] Yushuo Guan[2*]
Ruichuan An[1] Xinlong Chen[4] Yang Shi[1]
Pengfei Wan[2] Wentao Zhang[1] Yuanxing Zhang[2‡]

[1]Peking University [2]Kling Team, Kuaishou Technology
[3]Xi'an Jiaotong University [4]School of Artificial Intelligence, UCAS
libozhou@pku.edu.cn longo11070001@gmail.com

Abstract. The text encoder is a critical component of text-to-image and text-to-video diffusion models, fundamentally determining the semantic fidelity of the generated content. However, its development has been hindered by two major challenges: the lack of an efficient evaluation framework that reliably predicts downstream generation performance, and the difficulty of effectively adapting pretrained language models for visual synthesis. To address these issues, we introduce GRAN-TED, a paradigm to Generate Robust, Aligned, and Nuanced Text Embeddings for Diffusion models. Our contribution is twofold. First, we propose TED-6K, a novel text-only benchmark that enables efficient and robust assessment of an encoder's representational quality without requiring costly end-to-end model training. We demonstrate that performance on TED-6K, standardized via a lightweight, unified adapter, strongly correlates with an encoder's effectiveness in downstream generation tasks. Notably, under our experimental setup, compared with training a diffusion model from scratch, evaluating with TED-6K is about **750× faster**. Second, guided by this validated framework, we develop a superior text encoder using a novel two-stage training paradigm. This process involves an initial fine-tuning stage on a Multimodal Large Language Model for better visual representation, followed by a layer-wise weighting method to extract more nuanced and potent text features. Our experiments show that the resulting GRAN-TED encoder not only achieves state-of-the-art performance on TED-6K but also leads to demonstrable performance gains in text-to-image and text-to-video generation.

1 Introduction

Diffusion models have recently emerged as the dominant paradigm for text-to-image (T2I) [3, 17, 27, 48, 51, 52] and text-to-video (T2V) [16, 25, 43, 58, 66] generation. Central to these models is the text encoder, which transforms textual

* Equal contribution.
† Work done during an internship at Kling Team.
‡ Corresponding author.

prompts into semantic representations that guide the visual synthesis process. An effective text encoder must accurately represent a wide range of semantic elements, including subject features, static attributes, spatial relationships, and temporal events. This capability is crucial for enabling the generative model to achieve compositional generalization—synthesizing novel scenes from combinations of concepts not explicitly seen during training. Despite its critical role, the importance of the text encoder was largely overlooked in the early stages of diffusion model development, with most research focusing on the diffusion mechanism itself. This oversight is consequential, as the quality of the text encoder fundamentally determines the semantic fidelity between the prompt and the generated visual content.

This trend has begun to reverse, with a clear architectural evolution in text encoders. The community has progressed from using early CLIP-based encoders [52] to more powerful models like T5 [27], and has now converged on integrating Large Language Models (LLMs) as the state-of-the-art backbone [17,59]. The rationale is clear: the advanced semantic understanding of LLMs promises to directly translate into higher-fidelity visual generation, improving the alignment between text and image.

However, despite the adoption of these powerful LLMs, generative models still exhibit significant challenges in prompt fidelity. Common failures observed in practical applications and on benchmark evaluations include incorrect object counts, misinterpretation of relational or referential phrases, attribute binding errors, and a failure to adhere to negative constraints. As the bridge between textual descriptions and visual synthesis, a more accurate and robust text representation is paramount to addressing these shortcomings. Developing a superior text encoder tailored for generative models presents two primary challenges:

- **Evaluating Representation Quality for Generation**: A significant hurdle is the lack of a suitable evaluation framework. Standard NLP benchmarks do not correlate directly with visual generation. Furthermore, information retrieval metrics not only miss compositional complexity and nuanced world knowledge, but their single-vector paradigm inherently misaligns with how contemporary diffusion models utilize text embeddings. Conversely, evaluating an encoder via end-to-end training of a full generative model is computationally prohibitive and inefficient.
- **Adapting LLM Representations for Visual Synthesis**: The second challenge lies in effectively adapting the rich features of a pre-trained LLM for the specific demands of visual generation. It is non-trivial to ensure that the semantic information encoded in the LLM's latent space is represented in a manner that is precise, unambiguous, and fully interpretable by the diffusion model, thereby translating improved textual understanding into tangible gains in generation fidelity.

To address these two challenges, we propose GRAN-TED to Generate Robust, Aligned, and Nuanced Text Embedding for Diffusion models. We first introduce a robust and unified evaluation framework centered on a novel text-only benchmark named TED-6K. The TED-6K benchmark, composed of visual

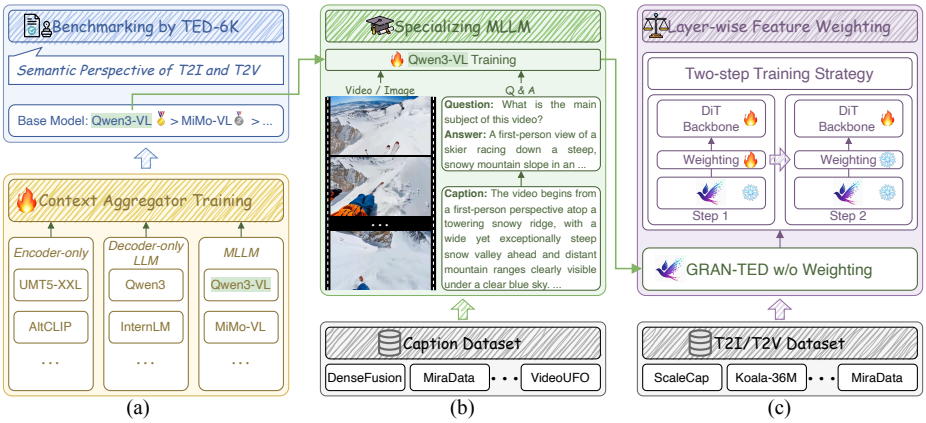

Fig. 1: An overview of our complete framework, integrating our evaluation and development pipelines. Figure (a) illustrates the TED-6K benchmark and an evaluation framework designed to assess the representational capabilities of text encoders. Within this framework, a context aggregator is utilized to aggregate sequential information. Figure (b) shows the construction of our TED Encoder, where Qwen3-VL-8B-Instruct is fine-tuned on a curated VQA and captioning dataset to specialize the MLLM. Figure (c) depicts our final GRAN-TED solution, which incorporates a learnable layer-wise weighting module to generate GRAN-TED for diffusion models.

captions and verified true/false statements, allows for efficient and robust assessment of an encoder's representational quality. We demonstrate that an encoder's performance on TED-6K, which is measured by representation similarity and standardized via a lightweight adapter, strongly correlates with its effectiveness in downstream text-to-image and text-to-video generation. In our experimental setup , training a T2I diffusion model from scratch for evaluation takes about 50 hours, whereas evaluating an encoder with TED-6K takes only about 4 minutes, enabling **750× faster** iteration.

Leveraging this validated framework, we then explore a two-stage paradigm to develop a superior text encoder GRAN-TED. In the first stage, we finetune Qwen3-VL-8B-Instruct [57] for better aligned latent visual representation towards generation. In the second stage, we implement a novel layer-wise weighting method, which extracts more nuanced text features from the layers of the trained model in the first stage. Our experiments confirm that this two-stage paradigm yields a new robust text encoder on TED-6K, simultaneously leading to demonstrable performance gains in T2I and T2V models.

2 Related Works

2.1 Text Encoder for Diffusion Models.

In diffusion models, the selection of text encoders is diverse. Early diffusion models commonly employ the text encoder from CLIP [49] or the encoder module

from T5 [50], since their pre-training objectives are inherently designed to efficiently encode text into semantic representations [47, 52]. As the potential of decoder-only LLMs [41, 69] for text encoding tasks is gradually unveiled [5, 21], recent research efforts have begun to select LLMs or MLLMs [2, 34, 57] as text encoders [31, 35, 42, 58] for diffusion models. Other studies explore deeper fusion between LLM representations and diffusion transformers [37, 56]. These works primarily study how to insert or combine strong language features inside a generator, while our focus is on efficiently evaluating and constructing the text encoder itself.

Besides leveraging the single feature layer from a single text model, some approaches employ feature fusion in the text encoder to enhance semantic capabilities. Seedream [17] integrates LLMs and a glyph-aligned model as the text encoder. Furthermore, although most contemporary diffusion models conventionally utilize hidden states from a single layer (typically the penultimate layer) of the text encoder as text conditioning, recent research [29, 59] demonstrates the efficacy of a superior feature aggregation strategy. They demonstrate that normalizing and averaging the hidden states from all layers (Norm-Avg) yields a stronger text representation than using any single-layer features or simple averaging (Avg):

$$c_{\text{text}} = \frac{1}{N} \sum_{i=1}^{N} \text{LayerNorm}(h_i) \tag{1}$$

However, these methods may not fully explore the connection between the different representational capabilities of each layer of the language model and downstream tasks. Our layer-wise weighting directly extends this line by learning the layer contribution and then freezing it to avoid a drifting conditioning distribution during diffusion training. Further studies also explore post-training strategies for the text encoder after the end-to-end training of diffusion models [7, 33], or distill large text encoders into smaller ones for efficient T2I systems [61]. These directions are complementary to GRAN-TED: post-training methods refine an already trained generator, while distillation changes deployment cost. In contrast, TED-6K provides a text-only way to compare candidate encoders before expensive generator training, and GRAN-TED specializes the encoder used for text-conditioned generation.

2.2 Text-alignment Benchmarks for Text Encoders.

Prevailing text-alignment benchmarks focus on assessing the correspondence between the final visual output and the input text prompt. For instance, GenAI-Bench [36] leverages a Vision-Language Model (VLM) to judge the alignment of generated images in terms of various attributes. Similarly, VIEScore [26] and GPT4-Eval [70] employ GPT-4V [46] to directly produce an image-text alignment score based on the visual output. These methods are costly and inefficient for encoder selection, since they assess the diffusion model as a whole, rather than isolating the text encoder's specific contribution. This necessitates the training

of a distinct diffusion model for each candidate encoder, which is prohibitively costly.

Some other methods leverage existing retrieval benchmarks and LLM benchmarks as proxy evaluations. Retrieval-based benchmarks like MTEB [44] are commonly-used proxy tasks. The text embeddings are encoded into a single pooled vector in these benchmarks, which has a granularity mismatch with the usage in diffusion models, since the text embeddings in diffusion models leverage the full sequence of hidden states for conditioning [16, 17, 25, 43, 48, 58]. LLM benchmarks [1, 8, 15, 20, 54, 65] are not suitable proxy tasks either, since their metrics focus on the perception and reasoning abilities, rather than the text-conditioning capability needed for visual generation. TED-6K differs from these benchmarks by directly testing whether text-condition representations distinguish fine-grained visual statements without training a full generator.

3 Method

3.1 Overview

As illustrated in Fig. 1, we have developed a robust pipeline designed to identify the optimal text encoder for diffusion models from a vast pool of candidates. First, we introduce TED-6K, a benchmark specifically tailored to the domain of visual generation. Unlike traditional methods [19, 26, 36], TED-6K serves as a text-only evaluation benchmark, enabling a multidimensional assessment of an encoder's representational capabilities without the prohibitive cost of training a full diffusion model end-to-end. Next, we develop a unified context aggregator to ensure a fair comparison across diverse architectures, including CLIP, T5, LLMs and MLLMs, which allows us to equitably evaluate the representational quality of various models on the TED-6K. Guided by this rigorous screening process, we identify Qwen3-VL-8B-Instruct [57] as the superior backbone. Building upon this backbone, we fine-tune it on the VQA dataset to develop an enhanced text encoder, termed GRAN-TED. Furthermore, we investigate strategies to efficiently fuse distinct feature layers within GRAN-TED, thereby maximizing the quality of the resulting text embeddings. This layer-wise fusion is the second construction step of GRAN-TED after MLLM specialization, rather than a separate diffusion-backbone contribution. The specifics of the benchmark formulation, the context aggregator architecture, our developed encoder, and the feature fusion strategy are detailed in Sec. 3.2 - Sec. 3.5, respectively.

3.2 Text-only Benchmark for Evaluation

We first build a text-only benchmark TED-6K illustrated in Fig. 2. Our goal is to systematically create a dataset capable of probing a text encoder's representational capabilities across key semantic dimensions. To this end, we design and execute the following multi-stage data pipeline:

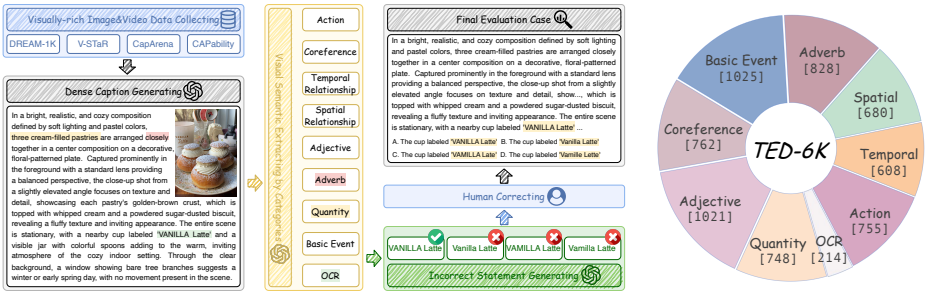

Fig. 2: *Left.* The data construction pipeline for TED-6K, consisting of four stages: (1) Data Curation and Filtering; (2) Base Caption Generation; (3) Semantic Pair Construction; (4) Human Verification. *Right.* The data Composition of the TED-6K dataset.

- **Raw Data Curation and Filtering:** We begin by collecting raw image and video from several high-quality open-source datasets, including DREAM1K [60], CAPability [39], V-STaR [11], and CapArena [10]. These sources are selected for their diversity, high visual fidelity, and semantic complexity.
- **Generation of Fine-Grained Base Captions:** For each filtered visual sample, we utilize the advanced capabilities of Gemini 2.5 Pro [13] to generate a fine-grained descriptive caption. We guide the model to produce descriptions of the utmost detail, while ensuring maximum coverage of our predefined semantic categories.
- **Construction of Multi-Dimensional Semantic Evaluation Pairs:** To probe specific semantic capabilities, we construct evaluation data for each base caption from eight predefined semantic perspectives: action, spatial relationship, temporal relationship, coreference, adjectives, adverbs, quantity, OCR, and basic event. First, if the original base caption contains information that could be articulated from a given perspective, we prompt the model to generate one positive statement consistent with its content. Subsequently, to construct high-quality negative samples, the model performs three distinct types of targeted semantic modifications on this positive statement, such as attribute swapping or relation reversal. A critical constraint is that each modification must create an explicit semantic contradiction with the original base caption, generating plausible yet factually incorrect statements.
- **Rigorous Human Verification:** Finally, a manual two-step verification is conducted: annotators first confirm each positive statement matches the original caption, then ensure negative statements are highly confusable with the positive ones. They explicitly target "hard-negatives" featuring high lexical similarity and subtle character variations.

This pipeline culminates in a dataset of 6,641 evaluation instances, each comprising a source caption, a corresponding positive statement, and a set of semantically contradictory negative statements. The specific distribution of evaluation instances across different semantic categories is illustrated in Fig. 2.

3.3 Sentence-level Context Aggregator

It is difficult to fairly evaluate the representation capability of different kinds of text encoders with traditional metrics, like single-vector retrieval or Question-Answer (QA) accuracy. Most text encoders like LLMs or MLLMs generally exhibit suboptimal single-vector retrieval performance; furthermore, this single-vector paradigm inherently misaligns with how contemporary diffusion models utilize text embeddings. Meanwhile, evaluating text encoders for diffusion models via question-answering (QA) accuracy is misleading. This approach is problematic because it cannot be applied to encoders like CLIP that lack QA functionality. More importantly, it conflates two distinct mechanisms within LLMs. QA relies on causal decoding using prefilling features from all model layers to infer an answer, whereas a Diffusion Transformer (DiT) typically uses a static embedding as its conditioning. Therefore, an LLM's reasoning ability in QA is not a reliable proxy for its ability to generate high-quality representations for visual generation.

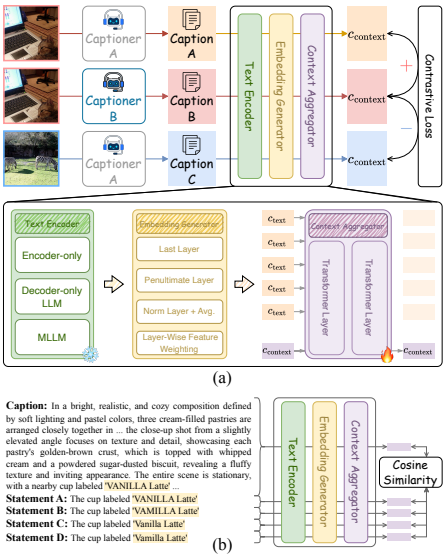

Fig. 3: The context aggregator architecture and its training&inference process.

To fairly assess the representation capability of different text encoders when used in DiTs, we design a unified context aggregator that mimics how DiTs consume text embeddings, and evaluate them at the representation level on TED-6K. The context aggregator, $\mathcal{A}_{\text{context}}$, consists of two self-attention Transformer layers and a learnable context token, c_{context}. It functions by aggregating the sentence-level semantics from an input text representation, c_{text}, into the context token. This information flow can be represented as:

$$c'_{\text{context}} \leftarrow \mathcal{A}_{\text{context}}([c_{\text{text}}, c_{\text{context}}]) \tag{2}$$

where c'_{context} is the updated context token, now infused with the information from c_{text}. The detailed architecture is provided in the Appendix A.

Fig. 3 illustrates the overall workflow of our evaluation framework. To construct the positive pairs for training, we employed the following strategy for each image: using the Qwen3-VL-235A22B-Instruct, we generated two semantically similar but textually diverse captions by prompting it with two distinct prompts at a temperature of 1.0. Overall, we collect 500k caption pairs from DenseFusion [32], ScaleCap [68], MiraData [22], Koala-36M [62], and Video-UFO [63].

Notably, the data in the TED-6K benchmark is distinct from our fine-tuning dataset, ensuring a fair evaluation of the model's generalization capabilities.

To enable the context aggregator $\mathcal{A}_{\text{context}}$ to effectively aggregate core sentence-level semantics from the sequence of hidden states, we train it using a contrastive loss [45]. First, both captions from a positive pair are passed through a frozen text encoder (using various feature extraction configurations like Last Layer or Norm-Avg) to obtain their respective textual representations, c^A_{text} and c^B_{text}. Then, the aggregator, $\mathcal{A}_{\text{context}}$, processes each of these representations to produce their corresponding context tokens, $c'_{\text{context, A}}$ and $c'_{\text{context, B}}$. Finally, we apply the contrastive loss, which is trained to pull together the context tokens from the same image while pushing apart those from different images.

During the evaluation on TED-6K, a source caption and its candidate statements are passed through a frozen text encoder, feature extractor and the context aggregator that were specifically trained for it, to obtain their respective context tokens. An evaluation instance is marked as correct if the similarity score between the source caption and the positive statement is the highest among all candidates. The final accuracy is the proportion of correctly identified instances.

3.4 Specializing MLLMs as Text Encoders for Visual Generation

Upon the context aggregator and TED-6K, we find that MLLMs exhibit a significant advantage over other models when used as text encoders for generative tasks, detailed in Sec. 4.1. This superiority is largely attributable to their multimodal training, which forces the text encoder to learn a semantic space that inherently aligns with visual concepts.

However, the training objectives of most existing MLLMs are geared towards developing general-purpose multimodal reasoning capabilities, with their training data often incorporating tasks like visual grounding and multimodal reasoning. These tasks primarily enhance a model's reasoning abilities, while its core function as a text encoder for generative tasks—the capacity to provide high-quality textual representations—is not explicitly optimized.

These analyses raise a natural question: can we develop an even more potent model by fine-tuning a strong MLLM backbone on tasks specifically tailored for visual generation? Motivated by this hypothesis, we introduce GRAN-TED, a text encoder specialized for visual generation, which we developed by building upon the Qwen3-VL-8B-Instruct model. Specifically, we adopt a data-centric approach [4] to enhance the model's capability on text embeddings. We have meticulously collected a large-scale dataset of images/videos and their corresponding semantically rich captions. Based on this collection, we construct a massive set of Visual Question Answering (VQA) pairs from multiple perspectives directly relevant to visual generation, such as object attributes, spatial relationships, and temporal order. By fine-tuning the Qwen3-VL-8B-Instruct model on this highly targeted VQA dataset, we develop GRAN-TED, a text encoder enhanced for diffusion models.

3.5 Layer-wise Feature Weighting

Recent empirical studies [59] and our TED-6K results in Sec. 4.1 show that multi-layer representations are more effective than relying on a single hidden layer. In particular, normalizing and averaging hidden states (Norm-Avg) significantly outperforms simple averaging (Avg). In pre-norm LLMs, where the norm of hidden states tends to increase with layer depth [24,30], the Norm-Avg operation not only serves to stabilize the numerical scale of the features but also functions as an implicit, inverse-norm weighting scheme.

This insight motivates our layer-wise weighting module: if a fixed, implicit weighting is beneficial, an explicit and learnable weighting mechanism can adapt the contribution of each layer to generation-oriented conditioning. The module first applies layer normalization to the hidden state h_i of each layer. It then assigns a set of learnable scalar weights w_i to these normalized representations, which are converted into normalized aggregation weights α_i via a softmax function:

$$\alpha_i = \frac{\exp(w_i)}{\sum_{j=1}^{L} \exp(w_j)} \tag{3}$$

The final fused representation c_{text} is then computed as the weighted sum of these normalized representations:

$$c_{\text{text}} = \sum_{i=1}^{L} \alpha_i \cdot \text{LayerNorm}(h_i) \tag{4}$$

This module allows the model to autonomously learn the importance of features from different hierarchical levels as required by the task, thereby constructing a more informative textual representation.

However, the training objective of a diffusion model evolves: early stages focus on denoising low-frequency global structures, while later stages refine high-frequency details. Given that different layers of an LLM capture distinct types of semantic information [14, 55], the optimal weighting of textual features for these stages would likely differ. Consequently, a naive continuously learnable weighting scheme can drift with the evolving denoising objective. This changes the text-conditioning distribution seen by the diffusion model during training and creates a non-stationary condition that can hinder stable convergence.

To mitigate this instability, we introduce a two-step training strategy. We first train the layer weights jointly with the diffusion model for an initial number of steps, allowing them to learn a generally effective compromise across layers. After that, we freeze these learned weights, thereby fixing the text-conditioning distribution and providing a stable, high-quality text representation for the remainder of the model's training. We provide detailed discussion on the two-step training strategy in the Appendix B.

4 Experiments

We evaluate our TED-6K benchmark across several mainstream text encoders. To validate the benchmark's effectiveness, namely whether its scores can predict

real-world performance in downstream generative tasks, we select a representative subset of these encoders and train corresponding T2I and T2V models for a formal correlation analysis. Both the TED-6K and T2I/T2V evaluations show that GRAN-TED has superior text representation capability. The detailed experimental setup can be seen in Appendix E.

4.1 TED-6K Results

Table 1: Performance comparison of different text encoders and feature extraction strategies on our text-only benchmark.

Model Type	Model Name	Last Layer	Penultimate Layer	Norm-Avg
Encoder-only	UMT5-XXL [12]	44.60	46.89	51.41
	AltCLIP [9]	47.15	46.59	51.86
	SigLIP-SO400M-384	46.74	47.43	50.35
	OpenAI/CLIP-ViT-L-336 [49]	46.42	43.52	45.13
LLM	Qwen3-8B-Instruct	53.62	53.00	55.77
	Qwen3-8B-Base	53.37	54.19	55.94
	Qwen3-4B-Instruct	53.64	53.58	55.25
	Qwen3-4B-Base	52.84	53.06	54.98
	Qwen3-4B-Thinking	52.90	54.10	54.86
	Qwen3-32B-Instruct	53.23	54.42	56.98
	Llama3-8B-Instruct [18]	54.75	55.22	55.93
	InternLM3-8B-Instruct [6]	55.10	55.40	55.46
MLLM	Qwen3-VL-8B-Instruct	55.37	54.25	56.81
	Qwen3-VL-8B-Thinking	55.67	55.40	56.51
	Qwen3-VL-4B-Instruct	54.03	54.92	55.20
	Qwen3-VL-32B-Instruct	54.60	55.91	57.24
	MiMo-VL-7B [67]	54.54	56.29	56.48
	Ovis-2.5-9B [40]	54.66	54.92	56.26
	Gemma3-4b-it [23]	54.52	53.80	54.59
	InternVL3-9B-Instruct [71]	55.68	55.68	55.59
	GRAN-TED w/o weighting	55.87	55.13	57.22
	GRAN-TED	-	-	**57.42**

Tab. 1 presents the evaluation results of various text encoder types on our text-only benchmark. Fine-tuned on our curated dataset, GRAN-TED not only outperforms all peer models but also remarkably closes the gap with 32B models. The added SigLIP and CLIP baselines remain below the LLM and MLLM groups under the same aggregator protocol, supporting the need to evaluate stronger language-model-based encoders for diffusion conditioning.

A detailed analysis of Tab. 1 reveals several clear and insightful trends:

– **Potential of Decoder-only Architectures:** The benchmark results on TED-6K unequivocally reveal that decoder-only LLMs exhibit significantly superior representational capabilities as text encoders compared to traditional encoder-only models. Notably, this performance gap does not stem

Table 2: Performance breakdown of Qwen3-VL and GRAN-TED across nine semantic sub-tasks (Norm-Avg).

Model	Quant.	Adj.	Coref.	Event	Adv.	Spat.	OCR	Temp.	Act.
Qwen3-VL	52.01	46.91	55.77	76.98	47.71	46.32	35.98	63.32	68.87
GRAN-TED	52.87	45.74	57.16	77.86	45.73	47.11	37.12	65.99	71.19
Δ	+0.86	-1.17	+1.39	+0.88	-1.98	+0.79	+1.14	+2.67	+2.32

solely from model size. For instance, the 5B-parameter UMT5-XXL encoder still performs considerably worse than the smaller 4B Qwen3-VL-Instruct.

- **Value of Multimodal Training:** A key and noteworthy finding is that MLLMs consistently and significantly outperform their LLM backbones on our text-only benchmark, despite the evaluation itself being entirely devoid of visual inputs. We believe this advantage stems from the multimodal training process, which compels the text encoder to learn how to more effectively encode visually-grounded concepts within its hidden states, thereby enabling the language model component to better leverage these representations for downstream tasks like VQA.
- **Ambiguous Effect of Instruction Tuning:** In contrast to the clear trends observed in multimodal settings, instruction tuning does not exhibit a consistent impact on the representational capabilities on TED-6K.
- **Impact of Thinking model** Interestingly, thinking models exhibit a paradoxical effect: while they enhance the representational quality of single-layer features, they conversely degrade performance when multi-layer features are aggregated via the Norm-Avg strategy. This likely stems from their distinct reasoning behavior during QA, which fundamentally alters their prefill hidden states and the resulting representations.
- **Effective Scaling Trends:** The scaling law on the model size differs significantly depending on the feature extraction strategy. When relying solely on single-layer features, we observed no clear scaling trend. However, when using the Norm-Avg strategy, the model's score on TED-6K exhibits a much clearer positive correlation with its parameter size.
- **Importance of Feature Aggregation:** Comparing single-layer features, the penultimate layer lacks a consistent advantage over the final layer. However, aggregating multi-layer features via the Norm-Avg strategy generally outperforms using any single layer on TED-6K. Since different layers capture distinct syntactic and semantic granularities, their aggregation typically yields a richer text representation and superior overall performance.

Tab. 2 presents a detailed performance breakdown of the models on the various sub-tasks of TED-6K. GRAN-TED outperforms the baseline model across all semantic dimensions except for Adjective and Adverb. Further analysis indicates that the model excels at capturing macro-level events (e.g., "Action") but still has room for improvement in dimensions requiring fine-grained detail (e.g., "Spatial Relationship", "OCR"), highlighting a clear direction for future work.

4.2 Correlation with Generation Performance

All reported metrics are higher-is-better. TED-6K reports accuracy: an instance is correct when the source caption is most similar to the human-verified positive statement among all candidate statements. For GenAI-Bench [28], we follow its VQA-style evaluation and use the judge model's "yes"-token score for each prompt-output pair, with logit bias restricting the answer space to yes/no tokens. UnifiedReward-2 [64] and VideoAlign [38] report reward-model scores for image-text and video-text alignment, respectively.

Table 3: Correlation between our benchmark scores and T2I generation performance. All evaluated encoders are instruction models, except for UMT5-XXL.

Text Encoder	**TED-6K**	**GenAI↑**	**UnifiedReward↑**
UMT5-XXL (last-layer)	44.60	54.67	2.834
Qwen2.5-VL-7B (last-layer)	53.05	71.15	2.995
Qwen3-8B (last-layer)	53.62	72.42	2.994
Gemma3-4b-it (last-layer)	54.52	71.56	2.999
InternVL3.5-8B-MPO (last-layer)	55.31	74.46	3.003
Qwen2.5-VL-7B (norm-avg)	55.99	75.51	2.997
Qwen3-VL-8B (norm-avg)	56.81	76.17	3.028
GRAN-TED	**57.42**	**77.41**	**3.035**
Pearson Corr. (r) w/ TED-6K		0.9923	0.9778
p-value		1.11×10^{-6}	2.69×10^{-5}

Table 4: Correlation between our benchmark scores and T2V generation performance.

Text Encoder	**TED-6K**	**GenAI↑**	**VideoAlign↑**
Qwen3-8B (last-layer)	53.62	65.13	0.767
Qwen3-VL-8B (last-layer)	55.37	70.59	0.967
InternVL3.5-8B-MPO (last-layer)	55.31	68.70	1.002
Qwen3-VL-8B (norm-avg)	56.81	77.94	1.288
GRAN-TED	**57.42**	**80.33**	**1.424**
Pearson Corr. (r) w/ TED-6K		0.9739	0.9866
p-value		5.05×10^{-3}	1.87×10^{-3}

To validate our text-only benchmark and investigate the consistency of its scores with downstream generative performance, we conduct a quantitative analysis by calculating the Pearson correlation coefficient between our benchmark scores and downstream evaluations. For these downstream metrics, we utilize GenAI-Bench [28] for both T2I and T2V tasks, alongside UnifiedReward-2 [64]

scores on DrawBench [53] for T2I, and VideoAlign [38] rewards on TA-hard [38] for T2V. The comprehensive correlation results for the T2I and T2V tasks are presented in Tab. 3 and Tab. 4, respectively. We observe that the scores from our text-only benchmark exhibit a statistically significant and strong positive correlation with the models' generative performance across all datasets. Furthermore, we include qualitative image examples in Appendix C and video examples in Appendix D.

4.3 Ablation Study

To validate the effectiveness of the proposed method, we further select Qwen3-VL-8B-Instruct as the base model for our experiments. The results can be seen in Tab. 5. As shown in the table, our two-step training strategy proposed in Sec. 3.5 significantly improves performance over the vanilla Norm-Avg, yet it only introduces a small number of learnable parameters equal to the number of Transformer layers in the LLM.

Notably, when the two-step strategy is omitted and only learnable weights are introduced with continuous training, performance slightly degrades. This shows that naive learnable weighting is not sufficient by itself: continuously updated layer weights change the text-conditioning distribution seen by the DiT, so the denoising network must adapt to a moving condition space while it is also learning the visual generation task. This observation aligns with the argument in Sec. 3.5 that a non-stationary text condition can interfere with the training process. We visualize the weights assigned to each layer at different training steps during the training progress. As illustrated in Fig. 4, the evolution of the layer weights reveals two key phenomena.

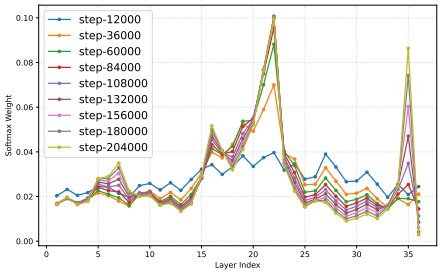

Fig. 4: Dynamics of the learnable layer weights over the course of continuous training (i.e., without the two-step strategy). The weight values shown are normalized via Softmax.

First, the weights of different layers exhibit non-monotonic dynamics. For instance, in the shallower regions of the model (e.g., layers 5 through 8), the weights for these layers collectively first increase and then decrease over the course of training. Second, the overall weight distribution displays persistent instability, failing to converge to an optimal solution. Even after 200k training steps, some deeper layers, such as the penultimate layer (-2), are still undergoing a substantial and consistent increase, while the weights of many other layers remain in flux. This observation provides a direct visual confirmation of our hypothesis that the text condition is indeed non-stationary under a contin-

uous training paradigm, thereby underscoring the necessity of our second-stage weight-freezing strategy.

Table 5: Ablation study over Qwen3-VL-8B-Instruct model. "Learnable Weights": trainable layer weights with continuous training. "Two-Step Training": the two-step training strategy to the learnable weights.

Method	GenAI-Bench	UnifiedReward
Norm-Avg (Fixed Weights)	76.17	3.028
+ Learnable Weights	75.94	3.021
+ Two-Step Training	**77.01**	**3.035**

Finally, we train diffusion models integrated with GRAN-TED, enhanced by both Norm-Avg and two-step layer-wise feature weighting. We then compare these models against a strong baseline trained on the Norm-Avg representations from the original Qwen3-VL-8B. The results demonstrate that models conditioned on our full methodology achieve superior generation performance across both modalities. For the T2I task (Table 3), GRAN-TED improves upon the baseline by +1.24 in GenAI and +0.007 in UnifiedReward. Similarly, for the T2V task (Table 4), it achieves gains of +2.39 in GenAI and +0.136 in VideoAlign. These downstream improvements are consistent with the +0.61 absolute increase in our TED-6K benchmark, underscoring the robust representational alignment capabilities of GRAN-TED.

4.4 Further Analysis

Performance Discrimination from TED-6K. The models achieve relatively similar scores as shown in Tab. 1. The reason is that part of the caption-statement pairs are simple for a comprehensive assessment, resulting in a low representational challenge. However, on the more demanding downstream tasks (Tab. 3-Tab. 4), which concentrate on difficult prompts, the performance gaps between models become significantly more pronounced. This validates that TED-6K is effectively discriminative.

Robustness of TED-6K. To ascertain the contextual dependency of the questions in TED-6K, we decouple the captions from their corresponding statements and shuffle them. The Qwen3-VL-8B-Instruct model is then tasked with direct question-answering on this shuffled data. Across five experimental runs, the model achieves an accuracy between 27.05% and 28.63%. This performance is approximately at the level of random guessing (25% for a four-option task), which demonstrates that the questions in TED-6K are sufficiently challenging and require the correct contextual caption to be answered accurately.

Aggregator vs. Single-vector Representation. A comparison between using a trained aggregator and applying mean pooling over the final layer's outputs reveals divergent results across architectures. According to Tab. 6, while

mean pooling improves representation quality for encoder-only models, it still falls short of the best performance on TED-6K and exhibits a trend inconsistent with downstream tasks. Conversely, for decoder-only LLMs, this method markedly impairs representational power. This justifies the necessity of our approach, which involves training an aggregator for effective model validation.

Defectiveness of QA-style Assessment. We compare the proposed similarity-based method with a direct QA method. For the QA task, we prompt the Qwen2.5-72B model to first convert each sample's statements into a question and then ask the text encoders to provide an open-ended answer. As detailed in Tab. 6, the model's powerful reasoning lead to uniformly accurate answers, which diminishes the method's ability to distinguish between test cases. This demonstrates the suitability of the similarity-based approach as a more discriminative evaluation metric.

Stability of TED-6K. To account for the variability from model training in our Aggregator approach, we train the aggregators for UMT5-XXL, Qwen3-8B-Instruct, and Qwen3-VL-8B-Instruct five times each. The maximum score variation across these runs is only 0.02 for any given model. This level of variance would not alter the assessed document order relations in the TED-6K benchmark, which confirms the stability of our evaluation methodology.

Table 6: Performance comparison of our aggregator-based method against mean-pooling approach and QA-style evaluations.

Model	Aggregator	Mean-pooling	Q& A
UMT5-XXL	44.65	50.64	97.81
Qwen3-8B-Instruct	53.62	39.18	97.90
Qwen3-VL-8B-Instruct	55.37	42.59	97.94

5 Conclusion

In this work, we address the critical role of the text encoder in T2I and T2V diffusion models. First, we establish an efficient evaluation framework centered on our novel text-only benchmark, TED-6K, for rapid encoder assessment. Second, we propose a two-stage paradigm to develop the GRAN-TED encoder, featuring targeted visual-semantic finetuning and layer-wise weighting. Experiments validate that GRAN-TED achieves state-of-the-art performance on TED-6K, translating into demonstrable improvements in semantic accuracy and compositional coherence for both T2I and T2V generation. The current scope of TED-6K is text-conditioned T2I/T2V generation. Extending the benchmark and training strategy to editing, TI2V, and 3D generation requires jointly evaluating text and visual conditions, and we leave these settings to future work.

References

1. AIME: Aime problems and solutions. `https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions` (2025), art of Problem Solving Wiki. Accessed: 2025-11-12 5
2. An, R., Yang, S., Lu, M., Zhang, R., Zeng, K., Luo, Y., Cao, J., Liang, H., Chen, Y., She, Q., et al.: Mc-llava: Multi-concept personalized vision-language model. arXiv preprint arXiv:2411.11706 (2024) 4
3. An, R., Yang, S., Zhang, R., Shen, Z., Lu, M., Dai, G., Liang, H., Guo, Z., Yan, S., Luo, Y., et al.: Unictokens: Boosting personalized understanding and generation via unified concept tokens. arXiv preprint arXiv:2505.14671 (2025) 1
4. Bai, T., Liang, H., Wan, B., Xu, Y., Li, X., Li, S., Yang, L., Li, B., Wang, Y., Cui, B., et al.: A survey of multimodal large language model from a data-centric perspective. arXiv preprint arXiv:2405.16640 (2024) 8
5. BehnamGhader, P., Adlakha, V., Mosbach, M., Bahdanau, D., Chapados, N., Reddy, S.: Llm2vec: Large language models are secretly powerful text encoders. arXiv preprint arXiv:2404.05961 (2024) 4
6. Cai, Z., Cao, M., Chen, H., Chen, K., Chen, K., Chen, X., Chen, X., Chen, Z., Chen, Z., Chu, P., et al.: Internlm2 technical report. arXiv preprint arXiv:2403.17297 (2024) 10
7. Chen, C., Wang, A., Wu, H., Liao, L., Sun, W., Yan, Q., Lin, W.: Enhancing diffusion models with text-encoder reinforcement learning. In: European Conference on Computer Vision. pp. 182–198. Springer (2024) 4
8. Chen, X., Lin, W., Hua, J., Yao, L., Ding, Y., Li, B., Zeng, B., Shi, Y., Liu, Q., Zhang, Y., et al.: Diadem: Advancing dialogue descriptions in audiovisual video captioning for multimodal large language models. arXiv preprint arXiv:2601.19267 (2026) 5
9. Chen, Z., Liu, G., Zhang, B.W., Yang, Q., Wu, L.: Altclip: Altering the language encoder in clip for extended language capabilities. In: Findings of the Association for Computational Linguistics: ACL 2023. pp. 8666–8682 (2023) 10
10. Cheng, K., Song, W., Fan, J., Ma, Z., Sun, Q., Xu, F., Yan, C., Chen, N., Zhang, J., Chen, J.: Caparena: Benchmarking and analyzing detailed image captioning in the llm era. arXiv preprint arXiv:2503.12329 (2025) 6
11. Cheng, Z., Hu, J., Liu, Z., Si, C., Li, W., Gong, S.: V-star: Benchmarking video-llms on video spatio-temporal reasoning. arXiv preprint arXiv:2503.11495 (2025) 6
12. Chung, H.W., Constant, N., Garcia, X., Roberts, A., Tay, Y., Narang, S., Firat, O.: Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining. arXiv preprint arXiv:2304.09151 (2023) 10
13. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025) 6
14. Dar, G., Geva, M., Gupta, A., Berant, J.: Analyzing transformers in embedding space. arXiv preprint arXiv:2209.02535 (2022) 9
15. Fu, Y., Li, B., Li, L., Zhang, W., Xie, T.: The first prompt counts the most! an evaluation of large language models on iterative example-based code generation. Proceedings of the ACM on Software Engineering **2**(ISSTA), 1583–1606 (2025) 5
16. Gao, Y., Guo, H., Hoang, T., Huang, W., Jiang, L., Kong, F., Li, H., Li, J., Li, L., Li, X., et al.: Seedance 1.0: Exploring the boundaries of video generation models. arXiv preprint arXiv:2506.09113 (2025) 1, 5

17. Gong, L., Hou, X., Li, F., Li, L., Lian, X., Liu, F., Liu, L., Liu, W., Lu, W., Shi, Y., et al.: Seedream 2.0: A native chinese-english bilingual image generation foundation model. arXiv preprint arXiv:2503.07703 (2025) 1, 2, 4, 5
18. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024) 10
19. Guo, Z., Chen, X., Zhang, R., An, R., Qi, Y., Jiang, D., Li, X., Zhang, M., Li, H., Heng, P.A.: Are video models ready as zero-shot reasoners? an empirical study with the mme-cof benchmark. arXiv preprint arXiv:2510.26802 (2025) 5
20. Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., Steinhardt, J.: Measuring massive multitask language understanding. arXiv preprint arXiv:2009.03300 (2020) 5
21. Jiang, Z., Meng, R., Yang, X., Yavuz, S., Zhou, Y., Chen, W.: Vlm2vec: Training vision-language models for massive multimodal embedding tasks. arXiv preprint arXiv:2410.05160 (2024) 4
22. Ju, X., Gao, Y., Zhang, Z., Yuan, Z., Wang, X., Zeng, A., Xiong, Y., Xu, Q., Shan, Y.: Miradata: A large-scale video dataset with long durations and structured captions (2024), https://arxiv.org/abs/2407.06358 7
23. Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., Rouillard, L., et al.: Gemma 3 technical report. arXiv preprint arXiv:2503.19786 **4** (2025) 10
24. Kim, J., Lee, B., Park, C., Oh, Y., Kim, B., Yoo, T., Shin, S., Han, D., Shin, J., Yoo, K.M.: Peri-ln: Revisiting normalization layer in the transformer architecture. arXiv preprint arXiv:2502.02732 (2025) 9
25. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024) 1, 5
26. Ku, M., Jiang, D., Wei, C., Yue, X., Chen, W.: Viescore: Towards explainable metrics for conditional image synthesis evaluation. arXiv preprint arXiv:2312.14867 (2023) 4, 5
27. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., Lacey, K., Levi, Y., Li, C., Lorenz, D., Müller, J., Podell, D., Rombach, R., Saini, H., Sauer, A., Smith, L.: Flux.1 kontext: Flow matching for in-context image generation and editing in latent space (2025), https://arxiv.org/abs/2506.15742 1, 2
28. Li, B., Lin, Z., Pathak, D., Li, J., Fei, Y., Wu, K., Ling, T., Xia, X., Zhang, P., Neubig, G., et al.: Genai-bench: Evaluating and improving compositional text-to-visual generation. arXiv preprint arXiv:2406.13743 (2024) 12
29. Li, B., Guan, Y., Li, H., Zeng, B., Ji, Y., Ding, Y., Wan, P., Gai, K., Zhang, Y., Zhang, W.: Semantic routing: Exploring multi-layer llm feature weighting for diffusion transformers. arXiv preprint arXiv:2602.03510 (2026) 4
30. Li, B., Xue, X., Yang, S., Shi, Y., Chen, X., Guan, Y., Zhang, Y., Zhang, W.: The unseen bias: How norm discrepancy in pre-norm mllms leads to visual information loss. arXiv preprint arXiv:2512.08374 (2025) 9
31. Li, P., Yu, P., Liu, Z., He, W., Pan, X., Rao, X., Wei, T., Chen, W.: Ldgen: Enhancing text-to-image synthesis via large language model-driven language representation. arXiv preprint arXiv:2502.18302 (2025) 4
32. Li, X., Zhang, F., Diao, H., Wang, Y., Wang, X., Duan, L.: Densefusion-1m: Merging vision experts for comprehensive multimodal perception. Advances in Neural Information Processing Systems **37**, 18535–18556 (2024) 7

33. Li, Y., Liu, X., Kag, A., Hu, J., Idelbayev, Y., Sagar, D., Wang, Y., Tulyakov, S., Ren, J.: Textcraftor: Your text encoder can be image quality controller. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7985–7995 (2024) 4
34. Lin, W., Wei, X., An, R., Gao, P., Zou, B., Luo, Y., Huang, S., Zhang, S., Li, H.: Draw-and-understand: Leveraging visual prompts to enable mllms to comprehend what you want. arXiv preprint arXiv:2403.20271 (2024) 4
35. Lin, W., Wei, X., An, R., Ren, T., Chen, T., Zhang, R., Guo, Z., Zhang, W., Zhang, L., Li, H.: Perceive anything: Recognize, explain, caption, and segment anything in images and videos (2025), https://arxiv.org/abs/2506.05302 4
36. Lin, Z., Pathak, D., Li, B., Li, J., Xia, X., Neubig, G., Zhang, P., Ramanan, D.: Evaluating text-to-visual generation with image-to-text generation. In: European Conference on Computer Vision. pp. 366–384. Springer (2024) 4, 5
37. Liu, B., Akhgari, E., Visheratin, A., Kamko, A., Xu, L., Shrirao, S., Lambert, C., Souza, J., Doshi, S., Li, D.: Playground v3: Improving text-to-image alignment with deep-fusion large language models. arXiv preprint arXiv:2409.10695 (2024) 4
38. Liu, J., Liu, G., Liang, J., Yuan, Z., Liu, X., Zheng, M., Wu, X., Wang, Q., Xia, M., Wang, X., et al.: Improving video generation with human feedback. arXiv preprint arXiv:2501.13918 (2025) 12, 13
39. Liu, Z., Xie, C.W., Wen, B., Yu, F., Chen, J., Zhang, B., Yang, N., Li, P., Li, Y., Gao, Z., et al.: What is a good caption? a comprehensive visual caption benchmark for evaluating both correctness and thoroughness. arXiv preprint arXiv:2502.14914 (2025) 6
40. Lu, S., Li, Y., Xia, Y., Hu, Y., Zhao, S., Ma, Y., Wei, Z., Li, Y., Duan, L., Zhao, J., et al.: Ovis2. 5 technical report. arXiv preprint arXiv:2508.11737 (2025) 10
41. Luo, Y., An, R., Zou, B., Tang, Y., Liu, J., Zhang, S.: Llm as dataset analyst: Subpopulation structure discovery with large language model. In: European Conference on Computer Vision. pp. 235–252. Springer (2024) 4
42. Ma, B., Zong, Z., Song, G., Li, H., Liu, Y.: Exploring the role of large language models in prompt encoding for diffusion models. arXiv preprint arXiv:2406.11831 (2024) 4
43. Ma, G., Huang, H., Yan, K., Chen, L., Duan, N., Yin, S., Wan, C., Ming, R., Song, X., Chen, X., et al.: Step-video-t2v technical report: The practice, challenges, and future of video foundation model. arXiv preprint arXiv:2502.10248 (2025) 1, 5
44. Muennighoff, N., Tazi, N., Magne, L., Reimers, N.: Mteb: Massive text embedding benchmark. arXiv preprint arXiv:2210.07316 (2022) 5
45. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018) 8
46. OpenAI: Gpt-4v(ision) system card (2024) 4
47. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023) 4
48. Qin, Q., Zhuo, L., Xin, Y., Du, R., Li, Z., Fu, B., Lu, Y., Yuan, J., Li, X., Liu, D., et al.: Lumina-image 2.0: A unified and efficient image generative framework. arXiv preprint arXiv:2503.21758 (2025) 1, 5
49. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021) 3, 10

50. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research **21**(140), 1–67 (2020) 4
51. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International conference on machine learning. pp. 8821–8831. Pmlr (2021) 1
52. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 1, 2, 4
53. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Lopes, R.G., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic text-to-image diffusion models with deep language understanding. ArXiv **abs/2205.11487** (2022), `https://api.semanticscholar.org/CorpusID:248986576` 13
54. Shi, Y., Dong, Y., Ding, Y., Wang, Y., Zhu, X., Zhou, S., Liu, W., Tian, H., Wang, R., Wang, H., et al.: Realunify: Do unified models truly benefit from unification? a comprehensive benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22488–22497 (2026) 5
55. Skean, O., Arefin, M.R., LeCun, Y., Shwartz-Ziv, R.: Does representation matter? exploring intermediate layers in large language models. arXiv preprint arXiv:2412.09563 (2024) 9
56. Tang, B., Zheng, B., Paul, S., Xie, S.: Exploring the deep fusion of large language models and diffusion transformers for text-to-image synthesis. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 28586–28595 (2025) 4
57. Team, Q.: Qwen3-vl: Sharper vision, deeper thought, broader action. `https://qwen.ai/blog?id=99f0335c4ad9ff6153e517418d48535ab6d8afef` (9 2025), accessed: 2025-09-23 3, 4, 5
58. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025) 1, 4, 5
59. Wang, A.Z., Ge, S., Karras, T., Liu, M.Y., Balaji, Y.: A comprehensive study of decoder-only llms for text-to-image generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 28575–28585 (2025) 2, 4, 9
60. Wang, J., Yuan, L., Zhang, Y., Sun, H.: Tarsier: Recipes for training and evaluating large video description models. arXiv preprint arXiv:2407.00634 (2024) 6
61. Wang, L., Liu, D., Liu, X., He, X.: Scaling down text encoders of text-to-image diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 18424–18433 (2025) 4
62. Wang, Q., Shi, Y., Ou, J., Chen, R., Lin, K., Wang, J., Jiang, B., Yang, H., Zheng, M., Tao, X., et al.: Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 8428–8437 (2025) 7
63. Wang, W., Yang, Y.: Videoufo: A million-scale user-focused dataset for text-to-video generation. arXiv preprint arXiv:2503.01739 (2025) 7
64. Wang, Y., Zang, Y., Li, H., Jin, C., Wang, J.: Unified reward model for multimodal understanding and generation. arXiv preprint arXiv:2503.05236 (2025) 12
65. Wang, Z., Yao, D., Li, B., Ma, D., Li, B., Li, Z., Xiong, F., Cui, B., Tang, L., Zhang, W.: Text2vectorsql: Towards a unified interface for vector search and sql queries. arXiv preprint arXiv:2506.23071 (2025) 5

66. Wiedemer, T., Li, Y., Vicol, P., Gu, S.S., Matarese, N., Swersky, K., Kim, B., Jaini, P., Geirhos, R.: Video models are zero-shot learners and reasoners. arXiv preprint arXiv:2509.20328 (2025) 1
67. Xiaomi, L., Team, C.: Mimo-vl technical report. arXiv preprint arXiv:2506.03569 (2025) 10
68. Xing, L., Huang, Q., Dong, X., Zhang, P., Zang, Y., Cao, Y., Li, J., Ding, S., Zhang, W., Yu, N., et al.: Scalecap: Inference-time scalable image captioning via dual-modality debiasing. arXiv preprint arXiv:2506.19848 (2025) 7
69. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025) 4
70. Zhang, X., Lu, Y., Wang, W., Yan, A., Yan, J., Qin, L., Wang, H., Yan, X., Wang, W.Y., Petzold, L.R.: Gpt-4v (ision) as a generalist evaluator for vision-language tasks. arXiv preprint arXiv:2311.01361 (2023) 4
71. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025) 10