Text based generation of other modalities

1. Grundidee

Moderne KI-Systeme können Texteingaben nutzen, um Inhalte in anderen Modalitäten zu erzeugen oder zu steuern.
Dafür wird Text in einen Steuervektor umgewandelt, der Inhalt, Stil und Struktur der Ausgabe bestimmt.

2. Text als Steuervektor

Ein Textprompt geht durch mehrere Schritte:

1. Text-Prompt (z. B. „a small cat on a bed“)
2. Tokenisierung des Textes
3. Umwandlung in ein Text-Embedding (hochdimensionaler Vektor)
4. Nutzung dieses Vektors als Steuervektor
5. Generatives Modell erzeugt Bild, Video oder Audio

Ein Prompt wird in einen hochdimensionalen Vektor übersetzt, der Information über:

- Objekte
- Eigenschaften
- Beziehungen
- Stil/Atmosphäre

enthält.

Dieser Vektor wirkt wie ein Steuersignal für die generative KI.

3. CLIP (OpenAI, 2021)

CLIP (Contrastive Language–Image Pretraining) verbindet Text und Bild in einem gemeinsamen semantischen Raum.

- Es ist trainiert auf ca.400 Mio. ungefilterten Bild-Text-Paaren
    - Viel mehr im Vergleich zu bisherigen Datensätzen (ImageNet ca.1,2 Mio. Bilder)
- Training mittels kontrastiven Lernens
- ermöglicht Zero-Shot-Lernen
- erzeugt nichts selbst, sondern bewertet, ob ein Bild zum Prompt passt

CLIP ist die semantische Grundlage moderner Text-zu-Image/Text/Video-Modelle.

4. DINO – Starke visuelle Features

DINO (Distillation with No Labels) ist ein self-supervised Vision Transformer, der ohne Text arbeitet.

Trainingsprinzip:

Teacher & Student Vision Transformers

- Das Modell sieht verschiedene Ausschnitte desselben Bildes
- Student soll die Representation des Teachers nachahmen
- Dadurch entstehen stabile visuelle Repräsentationen

erkennt automatisch:

- Objektkonturen
- Strukturen von Bildern

DINO liefert präzise visuelle Features, die CLIP ergänzen.

5. Steuerung durch Prompts

Nachdem Text und visuelle Inhalte repräsentiert sind, nutzt das generative Modell den Text über drei zentrale Mechanismen:

1. Cross-Attention:

- Das Modell bezieht sich bei jedem Generierungsschritt aktiv auf den Prompt
    - “Welche Wörter sind für dieses Detail relevant?”

2. Conditioning:

- Text moduliert interne Schichten
    - Beeinflusst Stil, Atmosphäre, Perspektive, Struktur

3. Guidance:

- Das Modell erzeugt:
    - eine Version mit Prompt
    - eine Version ohne Prompt
- Die Differenz zeigt den Einfluss des Textes und kann verstärkt oder abgeschwächt werden.

Text bestimmt Inhalt, Stil, Atmosphäre und Komposition.

6. Generative Modelle

- Diffusionsmodelle: Text steuert jeden Rauschreduktionsschritt
- Videomodelle: Text hält zeitliche Konsistenz
- Audiomodelle: Text bestimmt Struktur, Instrumente, Stil

7. Zusammenfassung

- Text wird zu einem Steuervektor übersetzt
- CLIP verbindet Text & Bild semantisch
- DINO ergänzt CLIP mit starken visuellen Strukturen
- Generative Modelle setzen den Steuervektor in Inhalte um
- Text bestimmt Inhalt, Stil, Atmosphäre und Struktur

Nutzer*innen haben vollständige Kontrolle über die Generierung

9. Handoutquellen (Empfehlungen für tieferes Verständnis):

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al. (2021).
Learning Transferable Visual Models From Natural Language Supervision. arXiv.
https://arxiv.org/abs/2103.00020
(Abgerufen am 27. Oktober 2025)

OpenAI. (2021).
CLIP: Connecting text and images.
https://openai.com/index/clip/
(Abgerufen am 27. Oktober 2025)

Computerphile. (2019).
How AI "Understands" Images (CLIP).
YouTube.
https://www.youtube.com/watch?v=KcSXcpluDe4
(Abgerufen am 05. November 2025)

Meta AI. (2023).
DINOv2: Advancing self-supervised learning for computer vision.
https://ai.meta.com/blog/dino-v2-computer-vision-self-supervised-learning/
(Abgerufen am 14. November 2025)

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., et al. (2023).
DINOv2: Learning robust visual features without supervision. arXiv.
https://arxiv.org/pdf/2304.07193
(Abgerufen am 14. November 2025)

Voxel51 (2024).
A history of CLIP model training data advances.
Voxel51 Blog.

https://voxel51.com/blog/a-history-of-clip-model-training-data-advances
(Abgerufen am 22. November 2025)